# Constrained optimization problems in stopped Markov decision processes (Development of the optimization theory for the dynamic systems and their applications)

AUTHOR(S):

Horiguchi, Masayuki

CITATION:

Horiguchi, Masayuki. Constrained optimization problems in stopped Markov decision processes (Development of the optimization theory for the dynamic systems and their applications). 数理解析研究所講究録 2002, 1263: 83-102

ISSUE DATE:

2002-05

URL:

http://hdl.handle.net/2433/42038

数理解析研究所講究録 1263 巻 2002 年 83-102



# 停止マルコフ決定過程における制約条件付き最適化問題

(Constrained optimization problems in stopped Markov decision processes)

千葉大学大学院自然科学研究科　堀口 正之 (Masayuki HORIGUCHI)
Graduate School of Science and Technology,
Chiba University

**abstract**

In this paper, we study three cases of constrained optimization problems in stopped Markov decision processes (MDPs). We introduces the concept of randomization into stopping structure of stopped MDPs, which makes it possible to solve the problem through the corresponding Mathematical Programming formulation in terms of occupation measures treated mainly by Borkar[8]. The optimization problem for the case of finite states and finite actions is considered over stopping time $\tau$ constrained so that $\mathbb{E}\tau \leqq \alpha$ for some fixed $\alpha > 0$. Analyzing the equivalent Mathematical Programming we prove the existence of an optimal constrained pair of policy and stopping time and gives the characterization of constrained optimal pairs. Subsequently, the results for one-constrained case are extended to the case of vector-valued terminal reward and multiple cost constraints, where a Pareto optimal pair of policy and stopping time is characterized by Mathematical Programming formulation and Lagrangian approaches. In the latter half of this paper, the dynamic programming approaches to the constrained MDPs with countable state and compact action spaces are studied. Introducing a randomized stationary stopping time, the existence of an optimal pair of stationary policy and stopping time is proved utilizing a Lagrange multiplier. Also, using the idea of the one-step look ahead(OLA, cf. Ross[31]) policy an optimal constrained pair is sought concretely.

## 0. Introduction

The constrained optimization problem for Markov decision processes (MDPs), which is called constrained MDPs, has been studied by many authors (e.g., Altman[1, 2, 4], Beutler and Ross[7], Borkar[8], Derman[10], Frid[12], Hordijk and Kallenberg [17] and Sennott[33, 34]). For solving a constrained MDPs, there are two methods as well-known, i.e., Linear Programming (LP) and Lagrangian approaches.

An LP approach was introduced by Derman and Klein[11] and Derman [10] and further developed by Kallenberg[25] and Hordijk and Kallenberg[17] in the case of finite states. This approach converts an original constrained problem to an certain equivalent LP whose decision variables correspond to the occupation measure. That is, the value of the original constrained problem is equal to the value of LP and there is one-to-one correspondence between the optimal policies of the original constrained problem and the solutions to the LP. The extension of an LP approach to the case of countable state MDPs was presented by Altman[1, 2, 3, 4].

On the other hand, a Lagrangian approach was introduced by Beutler and Ross[7] for the case of the average expected reward and one constraint. By the correspond parametric dynamic programming equation, Beutler and Ross[7] showed that there exists an optimal constrained stationary policy requiring randomization between two actions in at most one state under some ergodic conditions. This Lagrangian approach was used to reduce the problem to an unconstrained problem and to characterize the constrained optimal policy. This approach was generalized to the countable state case by Sennott [33, 34].

This paper is also concerned with an optimal stopping model with a stopping time constrained for a stochastic process which is first studied by Nachman[29] and Kennedy[26]. They have charac-

terized the constrained optimal stopping time by a Lagrangian approach.

In this paper, we treat with a combined model of MDPs and stopping problem, called stopped MDPs, which was first introduced by Furukawa and Iwamoto [15] and Hordijk[16] independently. Furukawa and Iwamoto[15] showed the existence of an optimal pair of policy and stopping time associated with some optimality criterions. Hordijk[16] has considered this model from a standpoint of potential theory introducing the Lyapunov function method for MDPs. Stopped MDPs was further developed by Iwamoto[22], Furukawa[13] and Rieder[30]. Rieder[30] treated with the non-stationary and unbounded model, in which several results obtained in (Furukawa and Iwamoto[15] and Hordijk[16]) were extended and completed. Also, the general utility treatment for stopped MDPs was studied by Kadota et al.[23, 24].

In this paper, we study constrained optimization problems in stopped MDPs as follows:

We introduces the concept of randomization into stopping structure of stopped MDPs, which makes it possible to solve the problem through the corresponding Mathematical Programming formulation in terms of occupation measures treated mainly by Borkar[8]. The optimization problem for the case of finite states and finite actions is considered over stopping time $\tau$ constrained so that $\mathbb{E}\tau \leqq \alpha$ for some fixed $\alpha > 0$. Analyzing the equivalent Mathematical Programming we prove the existence of an optimal constrained pair of policy and stopping time and gives the characterization of constrained optimal pairs. Subsequently, the results for one-constrained case are extended to the case of vector-valued terminal reward and multiple cost constraints, where a Pareto optimal pair of policy and stopping time is characterized by Mathematical Programming formulation and Lagrangian approaches. In the latter half of this paper, the dynamic programming approaches to the constrained MDPs with countable state and compact action spaces are studied. Introducing a randomized stationary stopping time, the existence of an optimal pair of stationary policy and stopping time is proved utilizing a Lagrange multiplier. Also, using the idea of the one-step look ahead(OLA, cf. Ross[31]) policy an optimal constrained pair is sought concretely.

In Section 1, after describing the model, relevant notations and definitions are given. To solve constrained optimization problems in this paper, randomized stopping time(RST) is introduced by enlarging a sample space (cf. Assaf and Samuel-Cahn[5], Chow et al.[9], Irle[21] and Kennedy[26]). Another representation of RST coined by Irle [21], called *F-representation*, is presented and several types of RSTs are defined. Also, the constrained optimization problems treated in this paper are given. Moreover, a sufficient class, which is a subclass of all pairs of policies and RSTs and is sufficiently rich so that a optimal pair exists in it, is given.

In the subsequent sections (Section 2–4), constrained optimization problems in stopped MDPs, which are studied in Horiguchi[18, 19] and Horiguchi, Kurano and Yasuda[20], are treated. Section 2 is devoted to consider the optimization problem for a stopped MDPs with finite states and actions over stopping times $\tau$ constrained so that $\mathbb{E}\tau \leqq \alpha$ for some fixed $\alpha > 0$. The problem is solved through randomization of stopping times and Mathematical Programming formulation by occupation measures. For the case of fixed entry time, Altman[2] has formed an equivalent infinite Linear Programming for the total cost criteria and by analyzing the corresponding LP formulation has shown that there exists an optimal constrained stationary policy. However, we follow a somewhat different approach by converting the original constrained problem to Mathematical Programming formulation (*parametric LP*), since the stopped Markov decision model is controlled over not only policies but also stopping times. Two types of occupation measures, running and stopped are treated, but stopped occupation measure is shown to be expressed by running one. The properties of the set of running occupation measures which is achieved by different classes of pairs of policies and RSTs are introduced. Analyzing the equivalent Mathematical Programming problem formulated by running occupation measures corresponding with sta-

tionary policies and RSTs, the existence of an optimal constrained pair of stationary policy and stopping time requiring randomization in at most one state is proved. Also, numerical example is given. In Section 3, a optimization problem for stopped MDPs with vector-valued terminal reward and multiple running cost constraints in the framework similar to Section 2 is considered. The optimality is defined by the concept of efficiency based on a pseudo-order preference relation $\preccurlyeq_K$ induced by a closed convex cone $K$ in $\mathbb{R}^p$. Then a Pareto optimization with respect to the pseudo-order $\preccurlyeq_K$ is considered(cf. Furukawa[14], Wakuta[37]). Applying the idea of occupation measures and using the scalarization technique for vector maximization problems we obtain the equivalent Mathematical Programming problem and show the existence of a Pareto optimal pair of stationary policy and stopping time requiring randomization in at most $k$ states, where $k$ is the number of constraints. Also, introducing a corresponding Lagrange function, the saddle-point statements for the constrained problem are given, whose results are applied to obtain a related parametric Mathematical Programming, by which the problem is solved. Numerical examples are given to illustrate the results. In Section 4, the constrained optimization problem similar to the formulation treated in Section 2 is considered except that the model consists of countable state space and compact metric action space. In this section, the problem formulation is referred to Hordijk[16]. The problem is handled by solving a parametric dynamic programming equation produced from a Lagrangian approach. The concept of a randomized stationary stopping time, which is a mixed extension of the entry time of a stopping region, is introduced in order to prove the existence of an optimal constrained pair of stationary policy and stopping time. The proof is executed by applying a Lagrange multiplier method developed by Frid[12], Beutler and Ross[7] and Sennott[34]. Also, using the idea of the OLA policy an optimal constrained pair is derived concretely. The constrained Markov deteriorating system is illustrated as an example.

# 1 Stopped Markov decision processes

## 1.1 Stopped Markov decision processes

Let $S$ and $A$ be the finite sets denoted by $S = \{1, 2, \ldots, N_1\}$ and $A = \{1, 2, \ldots, N_2\}$. The stopped Markov decision model consists of five objects:

$$(S, A, \{p_{ij}(a) : i, j \in S, a \in A\}, c, r) \tag{1.1}$$

where $S$ and $A$ denote the state and action spaces respectively and $\{p_{ij}(a)\}$ is the law of motion, i.e., for each $(i, a) \in S \times A, p_{ij}(a) \geqq 0$ and $\sum_{j \in S} p_{ij}(a) = 1$ and $c = c(i, a)$ is a running cost function on $S \times A$ and $r = r(i)$ is a terminal reward function on $S$ when selecting "stop" in state $i$. When the system is in state $i \in S$, if we select "stop" the process terminates with the terminal reward $r(i)$. If we select "continue" and take an action $a \in A$, we move to a new state $j \in S$ selected according to the probability distribution $p_{i\cdot}(a)$ and the cost $c(i, a)$ is incurred. This process is repeated from the new state $j \in S$.

Similarly, another control model formulated with vector-valued terminal reward and multiple running costs is given as follows:

$$(S, A, \{p_{ij}(a) : i, j \in S, a \in A\}, \{c^l, l = 1, 2, \ldots, k\}, \boldsymbol{r}) \tag{1.2}$$

where $c^l = c^l(i, a)$, $l = 1, 2, \ldots, k$, are running cost functions on $S \times A$, which will be related to $k$ constraints, and $\boldsymbol{r} = \boldsymbol{r}(i) = (r^1(i), \ldots, r^p(i))$ is a vector-valued terminal reward function on $S$ when selecting "stop" in state $i$.

Let $x_t, a_t$ be the state and action at time $t$ and $h_t = (x_1, a_1, \ldots, x_t) \in (S \times A)^{t-1} \times S$ the history up to time $t (t \geqq 1)$. A policy for a controlling the system is a sequence $\pi = (\pi_1, \pi_2, \ldots)$ such that, for each $t \geqq 1$, $\pi_t$ is a conditional probability measure on A given history $h_t$ with $\pi_t(A | x_1, a_1, \ldots, x_t) = 1$ for each $(x_1, a_1, \ldots, x_t) \in (S \times A)^{t-1} \times S$. Let $\Pi$ denotes the set of all policies. A policy $\pi = (\pi_1, \pi_2, \ldots)$ is a Markov policy if $\pi_t$ is a function of only $x_t$, i.e., $\pi_t(\cdot | x_1, a_1, \ldots, x_t) = \pi_t(\cdot | x_t)$ for all $(x_1, a_1, \ldots, x_t) \in (S \times A)^{t-1} \times S$. A Markov policy $\pi = (\pi_1, \pi_2, \ldots)$ is stationary if there exists a con-

ditional probability on $A$, $w(\cdot|i)$, given $i \in S$ such that $\pi_t(\cdot|x_t) = w(\cdot|x_t)$ for all $x_t \in S$ and $t \geqq 1$, and denoted by $w^\infty = (w, w, \ldots)$, or simply by $w$. A stationary policy $w$ is called deterministic if there exists a map $h : S \to A$ with $w(h(i)|i) = 1$ for all $i \in S$ and such a policy is identified by $h$. The sets of all Markov, stationary and deterministic policies will be denoted by $\Pi_M, \Pi_S$ and $\Pi_D$ respectively. Note that $\Pi_D \subset \Pi_S \subset \Pi_M \subset \Pi$. The sample spaces is the product space $\Omega = (S \times A)^\infty$. Let $X_t$, $\Delta_t$ be random quantities such that $X_t(\omega) = x_t$ and $\Delta_t(\omega) = a_t$ for all $\omega = (x_1, a_1, x_2, a_2, \ldots) \in \Omega$. For any given policy $\pi \in \Pi$ and initial distribution $\beta$ on $S$ we can specify the probability measure $\mathbb{P}^\pi_\beta$ on $\Omega$ in a usual way.

Let $H_t = (X_1, \Delta_1, \ldots, X_t)$. We denote by $\mathcal{B}(H_t)$ the $\sigma$-field induced by $H_t$. Let $\mathcal{F}_t = \mathcal{B}(H_t), (t \geqq 1)$ and $\mathcal{F}_\infty$ be the smallest $\sigma$-field containing each $\mathcal{F}_t, t \geqq 1$. Let $\overline{N} = \{1, 2, \ldots\} \cup \{\infty\}$. We call a map $\tau : \Omega \to \overline{N}$ a stopping time w.r.t. the filtration $\mathcal{F} = \{\mathcal{F}_t, t \in \overline{N}\}$ if $\{\tau = t\} \in \mathcal{F}_t$ for all $t \in \overline{N}$. In order to solve our problems described in the sequel, we need to introduce randomized stopping time (cf. Chow et al.[9] and Kennedy[26]). To this purpose, enlarging $\Omega$ to $\overline{\Omega} := \Omega \times [0, 1]$, we can embed $(\Omega, \mathcal{F}_\infty)$ to $(\overline{\Omega}, \mathcal{F}_\infty \times \mathbb{B}_1)$, where $\mathbb{B}_1$ is Borel subsets of $[0, 1]$. For a filtration $\mathcal{F}^* = \{\mathcal{F}^*_t, t \in \overline{N}\}$ with $\mathcal{F}^*_t = \mathcal{F}_t \times \mathbb{B}_1$ we can assume without loss of generality that for each $t \in \overline{N}$

$$\mathcal{F}_t \subset \mathcal{F}^*_t. \tag{1.3}$$

We call a map $\overline{\tau} : \overline{\Omega} \to \overline{N}$ a randomized stopping time (hereafter called RST) w.r.t. $\mathcal{F}^*$ if $\{\overline{\tau} = t\} \in \mathcal{F}^*_t$ for each $t \in \overline{N}$. For simplicity, the upper bar of RST $\overline{\tau}$ will be omitted and written by $\tau$ with some abuse of notation. The class of RSTs w.r.t. $\mathcal{F}^*$ will be denoted by $\mathcal{S}$. For each initial distribution $\beta$ and each policy $\pi \in \Pi$, we denote the probability measure on $\overline{\Omega}$ by $\overline{\mathbb{P}}^\pi_\beta$, where $\overline{\mathbb{P}}^\pi_\beta = \mathbb{P}^\pi_\beta \times \lambda$ and $\lambda$ is Lebesgue measure on $\mathbb{B}_1$.

## 1.2 $F$-representation of RSTs

In this section, $F$-representation of RSTs given by Irle[21] will be extended to the case of the decision process considered in this paper by which Markov or stationary RSTs are defined.

For any RST $\tau \in \mathcal{S}$ and $t \in \overline{N}$, let $g_t(\omega) := \lambda(\{\tau = t\}_\omega)$ $(\omega \in \Omega)$, where $\{\tau = t\}_\omega$ is the $\omega$-section defined by $\{\tau = t\}_\omega = \{x \in [0, 1] | (\omega, x) \in \{\tau = t\}\}$. Note that $g_t$ is $\mathcal{F}_t$-measurable for $t \geqq 1$. From this $g_t$ $(t \in \overline{N})$, we define the set $f = (f_t)_{t \in \overline{N}}$ as follows:

$$f_t := \frac{g_t}{1 - \sum_{k=1}^{t-1} g_k}, \quad t \in \overline{N} \tag{1.4}$$

where if the denominator is 0 in (1.4) let $f_t = 1$.

Let $F = \{a = (a_j)_{j \in \overline{N}} : 0 \leqq a_j \leqq 1, a_\infty = 1$ and if $a_j = 1$ $a_i = 1$ for $i > j\}$. Then we have the following lemma.

**Lemma 1.2.1.**

(i) *$f : \Omega \to F$ and for each $t \in \overline{N}$ $f_t$ is $\mathcal{F}_t$-measurable.*

(ii) *For any initial distribution $\beta$ and pair $(\pi, \tau) \in \Pi \times \mathcal{S}$ and $t \in \overline{N}$,*

$$f_t = \frac{\overline{\mathbb{P}}^\pi_\beta(\tau = t|H_t)}{\overline{\mathbb{P}}^\pi_\beta(\tau \geqq t|H_t)}, \ \mathbb{P}^\pi_\beta\text{-a.s.} \tag{1.5}$$

(iii) *For any initial distribution $\beta$ and pair $(\pi, \tau) \in \Pi \times \mathcal{S}$,*

$$\begin{aligned} &\overline{\mathbb{E}}^\pi_\beta[\sum_{t=1}^{\tau-1} c(X_t, \Delta_t) + r(X_\tau)] \\ &= \sum_{t=1}^{\infty} (\mathbb{E}^\pi_\beta((1 - f_1) \cdots (1 - f_{t-1}) f_t \cdot \\ &\quad (\sum_{k=1}^{t-1} c(X_k, \Delta_k) + r(X_t)))). \end{aligned} \tag{1.6}$$

The set $f = (f_t)_{t \in \overline{N}}$ constructed from $\tau \in \mathcal{S}$ is called $F$-representation of $\tau$, denoted by $f^\tau = (f^\tau_t)_{t \in \overline{N}}$.

Let $f = (f_t)_{t \in \overline{N}}$ be any function $f : \Omega \to F$ such that for each $t \in \overline{N}$ $f_t$ is $\mathcal{F}_t$-measurable. From this $f$, we define $\tau^f : \Omega \times [0, 1] \to \overline{N}$ by

$$\tau^f(\omega, x) := \begin{cases} t \ \text{ for } \ x \in [\sum_{k=1}^{t-1} \overline{g}_k(\omega), \sum_{k=1}^{t} \overline{g}_k(\omega)), \\ \infty \ \text{ for } \ x \in [\sum_{k=1}^{\infty} \overline{g}_k(\omega), 1] \end{cases} \tag{1.7}$$

where

$$\overline{g}_t := (1 - f_1) \cdots (1 - f_{t-1}) f_t, \quad t \geqq 1. \tag{1.8}$$

Then, we have:

**Lemma 1.2.2.** (i) $\tau^f$ *is a RST w.r.t.* $\mathcal{F}^* = \{\mathcal{F}_t^*, t \in \overline{N}\}$.

(ii) $\tau^f$ *satisfies* (ii) *and* (iii) *of Lemma 1.2.1.*

Note that Lemma 1.2.1 and 1.2.2 show there is one-to-one correspondence between $\mathcal{S}$ and the set of $F$-representations $f = (f_t)_{t\in\overline{N}}$. Using this fact, we define several types of RSTs. Let $\tau \in \mathcal{S}$. For the corresponding $F$-representation $f^\tau = (f_t^\tau)_{t\in\overline{N}}$, by Lemma 1.2.1, $f_t^\tau$ is $\mathcal{F}_t$-measurable ($t \geqq 1$). So, $f_t^\tau$ is a function of $H_t = (X_1, \Delta_1, \ldots, X_t)$.

**Definition 1.** If $f_t^\tau$ is depending only on $X_t$, that is, $f_t^\tau(H_t) = f_t^\tau(X_t)$ for all $t \geqq 1$, the RST $\tau$ is called *Markov*. A Markov RST is called *stationary* if there exists a function $\delta : S \to [0,1]$ such that $f_t^\tau(X_t) = \delta(X_t)$ for all $t \geqq 1$, and denoted by $\delta^\infty$. When $\delta(i) \in \{0,1\}$ for all $i \in S$, the stationary RST $\delta^\infty$ is called *deterministic*.

We denote the sets of all Markov RSTs, all stationary RSTs and all deterministic RSTs by $\mathcal{S}_M, \mathcal{S}_S$ and $\mathcal{S}_D$ respectively.

## 1.3 Constrained optimization problems

For any $\alpha > 0$ and initial distribution $\beta$ on $S$, let

$$\Lambda(\alpha,\beta) := \{(\pi,\tau) \in \Pi \times \mathcal{S} \mid \overline{\mathbb{E}}_\beta^\pi \tau \leqq \alpha\} \qquad (1.9)$$

where $\overline{\mathbb{E}}_\beta^\pi$ is the expectation w.r.t. $\overline{\mathbb{P}}_\beta^\pi$. The pair belonging to $\Lambda(\alpha,\beta)$ will be called a constrained one. In Section 2 and 4, we will consider the constrained optimization problem(**COP**):

$$\textbf{COP}: \text{ Maximize } \overline{\mathbb{E}}_\beta^\pi[\sum_{t=1}^{\tau-1} c(X_t,\Delta_t) + r(X_\tau)]$$
$$\text{subject to } (\pi,\tau) \in \Lambda(\alpha,\beta).$$

On the other hand, in Section 3, we consider the vector-valued optimization problem with multiple constraints as follows.

For any $\alpha = (\alpha^1, \ldots, \alpha^k) \in \mathbb{R}^k$ and initial distribution $\beta$ on $S$, let $\Lambda^k(\alpha,\beta) := \{(\pi,\tau) \in \Pi \times \mathcal{S} \mid \overline{\mathbb{E}}_\beta^\pi \sum_{t=1}^{\tau-1} c^l(X_t,\Delta_t) \leqq \alpha^l \text{ for } l = 1,2,\ldots,k\}$. We shall define the vector-valued constrained optimization problem(**VCOP**):

$$\textbf{VCOP}: \text{ Maximize } \overline{\mathbb{E}}_\beta^\pi \boldsymbol{r}(X_\tau) := (\overline{\mathbb{E}}_\beta^\pi r^1(X_\tau), \ldots, \overline{\mathbb{E}}_\beta^\pi r^p(X_\tau))$$
$$\text{subject to } (\pi,\tau) \in \Lambda^k(\alpha,\beta).$$

## 1.4 Markov policies and Markov RSTs

In the following, we say that the set of $\Pi_M \times \mathcal{S}_M$ is a sufficient class to our optimization problems.

**Lemma 1.4.1.** *For any pair* $(\pi,\tau) \in \Pi \times \mathcal{S}$, *there exist a pair* $(v,\sigma) \in \Pi_M \times \mathcal{S}_M$ *such that*

$$\overline{\mathbb{P}}_\beta^\pi(X_t = i, \Delta_t = a, \tau > t) = \overline{\mathbb{P}}_\beta^v(X_t = i, \Delta_t = a, \sigma > t) \qquad (1.10)$$

*for* $i \in S, a \in A$.

# 2 Finite MDPs with a constraint([18])

## 2.1 One-constrained problem

In this section, we will consider the stopped Markov decision model

$$(S, A, \{p_{ij}(a) : i,j \in S, a \in A\}, c, r)$$

introduced in (1.1) where $S$ and $A$ be finite sets denoted by $S = \{1,2,\ldots,N_1\}$ and $A = \{1,2,\ldots,N_2\}$ and the constrained optimization problem as follows:

$$\textbf{COP}: \text{ Maximize } J(\beta,\pi,\tau) :=$$
$$\overline{\mathbb{E}}_\beta^\pi[\sum_{t=1}^{\tau-1} c(X_t,\Delta_t) + r(X_\tau)]$$
$$\text{subject to } (\pi,\tau) \in \Lambda(\alpha,\beta).$$

where $\Lambda(\alpha,\beta)$ is defined in (1.9).

The constrained pair $(\pi^*,\tau^*) \in \Lambda(\alpha,\beta)$ is called optimal if

$$J(\beta,\pi,\tau) \leqq J(\beta,\pi^*,\tau^*) \text{ for all } (\pi,\tau) \in \Lambda(\alpha,\beta).$$

## 2.2 Running and stopped occupation measures

We introduce, in this section, two types of occupation measures and consider the properties of them. Also, we formulate the Mathematical Programming problem which is proved to be equivalent to **COP**.

**Definition 2.** For any initial distribution $\beta$ and a pair $(\pi,\tau)$ with $\overline{\mathbb{E}}^{\pi}_{\beta}[\tau] < \infty$, we define the measure $x(\beta,\pi,\tau)$ on $S \times A$, called the *running occupation measure*, by

$$x(\beta,\pi,\tau;i,a) := \sum_{t=1}^{\infty} \overline{\mathbb{P}}^{\pi}_{\beta}(X_t = i, \Delta_t = a, \tau > t) \tag{2.1}$$

for $i \in S, a \in A$.

**Definition 3.** For any initial distribution $\beta$ and a pair $(\pi,\tau)$ with $\overline{\mathbb{E}}^{\pi}_{\beta}[\tau] < \infty$, we define the measure $y(\beta,\pi,\tau)$ on $S \times A$, called the *stopped occupation measure*, by

$$y(\beta,\pi,\tau;i,a) := \sum_{t=1}^{\infty} \overline{\mathbb{P}}^{\pi}_{\beta}(X_t = i, \Delta_t = a, \tau = t), \tag{2.2}$$

for $i \in S, a \in A$.

The *state* running and stopped occupation measures will be defined by $x(\beta,\pi,\tau;i) := \sum_{a\in A} x(\beta,\pi,\tau;i,a)$ and $y(\beta,\pi,\tau;i) := \sum_{a\in A} y(\beta,\pi,\tau;i,a)$ for all $i \in S$ respectively. Then, in the following lemma, the state stopped occupation measure is proved to be represented by the running one.

**Lemma 2.2.1.** *For any initial distribution $\beta$ and pair $(\pi,\tau) \in \Pi \times \mathcal{S}$ with $\overline{\mathbb{E}}^{\pi}_{\beta}[\tau] < \infty$ we have the following:*

(i) $x(\beta,\pi,\tau;i) < \infty$ *and* $y(\beta,\pi,\tau;i) < \infty$ *for all* $i \in S$.

(ii) $\overline{\mathbb{E}}^{\pi}_{\beta}[\tau] = \sum_{i\in S} x(\beta,\pi,\tau;i) + 1$.

(iii) $y(\beta,\pi,\tau;i) = \beta(i) + \sum_{j\in S,a\in A} x(\beta,\pi,\tau;j,a)p_{ji}(a) - x(\beta,\pi,\tau;i)$ *for all* $i \in S$.

For any $\delta : S \to [0,1]$ and conditional distribution $w(\cdot|i)$ on $A$ given $i \in S$, we define by $P^{\delta}(w)$ the $N_1 \times N_1$ matrix where $(i,j)$th element is $\sum_{a\in A} p_{ij}(a)w(a|i)(1-\delta(j)) := p_{ij}(w)(1-\delta(j))$ or simply $(P^{\delta}(w))_{ij}$. Let $\mathbb{R}^{N_1}$ be the set of real $N_1$-dimensional row vectors. With some abuse of notation, for any initial distribution $\beta$ and $(\pi,\tau) \in \Pi\times\mathcal{S}$, the row vector $x(\beta,\pi,\tau) \in \mathbb{R}^{N_1}$ is defined by

$$x(\beta,\pi,\tau) := (x(\beta,\pi,\tau;1),\ldots,x(\beta,\pi,\tau;N_1)).$$

If the distribution $\beta$ on $S$ is degenerate as $i \in S$, it is simply denoted by $i$.

**Lemma 2.2.2.** *Let $(w,\tau) \in \Pi_S \times \mathcal{S}_S$ with $\overline{\mathbb{E}}^{w}_{i}(\tau) < \infty$ for all $i \in S$. Then the state running occupation measure $x(\beta,w,\tau)$ is the unique solution to*

$$x = \beta(1-\delta) + xP^{\delta}(w),\ x \in \mathbb{R}^{N_1} \tag{2.3}$$

*where $\beta(1-\delta)$ is in $\mathbb{R}^{N_1}$ whose i-th component is $\beta(i)(1-\delta(i))$ and $\delta := f^{\tau} : S \to [0,1]$ is F-representation of $\tau$.*

Next, we present that the objective function $J(\beta,\pi,\tau)$ of **COP** is written by running and stopped occupation measures.

**Lemma 2.2.3.** *For $(\pi,\tau) \in \Pi\times\mathcal{S}$ with $\overline{\mathbb{E}}^{\pi}_{\beta}[\tau] < \infty$, we have*

$$J(\beta,\pi,\tau) = \sum_{i\in S,a\in A} c(i,a)x(\beta,\pi,\tau;i,a) + \sum_{i\in S} r(i)y(\beta,\pi,\tau;i). \tag{2.4}$$

Let $\mathbb{R}^{N_1\times N_2}$ be the set of real $N_1 \times N_2$ matrices. For any subset $U \subset \Pi \times \mathcal{S}$, let

$$\mathbf{X}^{\beta}_{\{\leqq\}\alpha}(U) = \{x(\beta,\pi,\tau;i,a)_{i\in S,a\in A} : (\pi,\tau) \in U,\ \overline{\mathbb{E}}^{\pi}_{\beta}[\tau] \leqq \alpha\}. \tag{2.5}$$

Note that $\mathbf{X}^{\beta}_{\{\leqq\}\alpha}(U) \subset \mathbb{R}^{N_1\times N_2}$. We introduce the Mathematical Programming(**MP(I)**) as follows.

**MP(I):** Maximize $\sum_{i\in S,a\in A} c(i,a)x(i,a) + \sum_{i\in S} r(i)y(i)$

subject to $x \in \mathbf{X}^{\beta}_{\{\leqq\}\alpha}(\Pi\times\mathcal{S})$, $y \in \mathbb{R}^{N_1}$ and

$$y(i) = \beta(i) + \sum_{j\in S,a\in A} x(j,a)p_{ji}(a) - x(i),$$

$i \in S$, where $x(i) = \sum_{a\in A} x(i,a)$.

Then, we have the following theorem whose proof follows easily from Lemma 2.2.3.

**Theorem 2.2.1.** **COP** *is equivalent to* **MP(I)**, *i.e., a pair $(\pi^*,\tau^*)$ is optimal for* **COP** *if and only if the corresponding $\{x(\beta,\pi^*,\tau^*;i,a)\} \in \mathbf{X}^{\beta}_{\{\leqq\}\alpha}(\Pi\times\mathcal{S})$ is optimal for* **MP(I)**.

## 2.3 Mathematical Programming and optimal pair

In this section, we present another Mathematical Programming formulation by which **COP** is explicitly solved.

For any $U \subset \Pi \times \mathcal{S}$, let $\mathbb{X}^{\beta}_{\{=\}\alpha}(U)$ be the set of $\mathbb{X}^{\beta}_{\{\leqq\}\alpha}(U)$ which is defined by replacing $\overline{\mathbb{E}}^{\pi}_{\beta}[\tau] \leqq \alpha$ with $\overline{\mathbb{E}}^{\pi}_{\beta}[\tau] = \alpha$ in (2.5).

**Theorem 2.3.1.**

$$\mathbb{X}^{\beta}_{\{\leqq\}\alpha}(\Pi\times\mathcal{S}) = \mathbb{X}^{\beta}_{\{\leqq\}\alpha}(\Pi_M\times\mathcal{S}_M) = \mathbb{X}^{\beta}_{\{\leqq\}\alpha}(\Pi_S\times\mathcal{S}_S), \tag{2.6}$$

*and*

$$\mathbb{X}^{\beta}_{\{=\}\alpha}(\Pi\times\mathcal{S}) = \mathbb{X}^{\beta}_{\{=\}\alpha}(\Pi_M\times\mathcal{S}_M) = \mathbb{X}^{\beta}_{\{=\}\alpha}(\Pi_S\times\mathcal{S}_S). \tag{2.7}$$

*Proof.* It is sufficient to prove (2.7). From Lemma 1.4.1 the first equality of (2.7) is shown. To prove the second part, for any running occupation measure $\{x(\beta,\pi,\tau;i,a)\} \in \mathbb{X}^{\beta}_{\{=\}\alpha}(\Pi \times \mathcal{S})$, we define $w \in \Pi_S$ and $\sigma^{\delta} \in \mathcal{S}_S$ with $\delta = f^{\sigma}$ by the following:

$$w(a|i) := \frac{x(\beta,\pi,\tau;i,a)}{x(\beta,\pi,\tau;i)} \quad \text{for } i \in S \text{ and } a \in A, \tag{2.8}$$

$$1-\delta(i) := \frac{x(\beta,\pi,\tau;i)}{\sum_{t=1}^{\infty}\overline{\mathbb{P}}^{\pi}_{\beta}(X_t = i, \tau \geqq t)} \quad \text{for } i \in S. \tag{2.9}$$

We note that

$$\overline{\mathbb{P}}^{\pi}_{\beta}(X_t = i, \tau \geqq t) = \overline{\mathbb{P}}^{\pi}_{\beta}(X_t = i, \tau > t-1)$$
$$= \sum_{j\in S, a\in A}\overline{\mathbb{P}}^{\pi}_{\beta}(X_{t-1} = j, \Delta_{t-1} = a, \tau > t-1)p_{ji}(a).$$

So, we get from (2.9) and (2.8)

$$x(\beta,\pi,\tau;i) = (1-\delta(i))\sum_{t=1}^{\infty}\overline{\mathbb{P}}^{\pi}_{\beta}(X_t = i, \tau \geqq t)$$
$$= (1-\delta(i))(\beta(i) + \sum_{j\in S, a\in A} x(\beta,\pi,\tau;j,a)p_{ji}(a))$$
$$= (1-\delta(i))(\beta(i) + \sum_{j\in S} x(\beta,\pi,\tau;j)(\sum_{a\in A} p_{ji}(a)w(a|j)))$$
$$= (1-\delta(i))\beta(i) + \sum_{j\in S} x(\beta,\pi,\tau;j)(P^{\delta}(w))_{ji}.$$

Applying Lemma 2.2.2, we have

$$x(\beta,\pi,\tau;i) = x(\beta,w,\sigma^{\delta};i), \ i \in S,$$

as required.∎

In order to drive another Mathematical Programming formulation, we need the definition of several basic sets. For simplicity, we put $(x_{ia}) = \{x_{ia}\}_{i\in S, a\in A} \in \mathbb{R}^{N_1\times N_2}$ and $\delta = \{\delta(i)\}_{i\in S} \in \mathbb{R}^{N_1}$. With some abuse of notation, $x_i = \sum_{a\in A} x_{ia}$ for $(x_{ia}) \in \mathbb{R}^{N_1\times N_2}$. For any initial distribution $\beta$ on $S$ and $\alpha(>1)$, let

$$\hat{\mathbb{Q}}_{\{\leqq\}\alpha} := \left\{ \begin{array}{l} ((x_{ia}),\delta) \in \mathbb{R}^{N_1\times N_2}\times\mathbb{R}^{N_1}: \\ \text{(i) } x_i = \beta(i)(1-\delta(i)) \\ \quad + \sum_{j\in S, a\in A} x_{ja}p_{ji}(a)(1-\delta(i)), \ i \in S \\ \text{(ii) } 0 \leqq \delta(i) \leqq 1, i \in S \\ \text{(iii) } \sum_{i\in S, a\in A} x_{ia} \leqq \alpha - 1 \\ \text{(iv) } x_{ia} \geqq 0, \quad i \in S, a \in A \end{array} \right\} \tag{2.10}$$

Let

$$\mathbb{Q}_{\{\leqq\}\alpha} := \{(x_{ia}) \in \mathbb{R}^{N_1\times N_2} : ((x_{ia}),\delta) \in \hat{\mathbb{Q}}_{\{\leqq\}\alpha} \text{ for some } \delta\}. \tag{2.11}$$

We denote by $\hat{\mathbb{Q}}_{\{=\}\alpha}$ the subset of $\hat{\mathbb{Q}}_{\{\leqq\}\alpha}$ obtained replacing (iii) in (2.10) by $\sum_{i\in S, a\in A} x_{ia} = \alpha - 1$ and by $\mathbb{Q}_{\{=\}\alpha}$ the set defined in (2.11) replacing $\hat{\mathbb{Q}}_{\{\leqq\}\alpha}$ by $\hat{\mathbb{Q}}_{\{=\}\alpha}$.

**Lemma 2.3.1.** *Both $\mathbb{Q}_{\{\leqq\}\alpha}$ and $\mathbb{Q}_{\{=\}\alpha}$ are compact and convex.*

*Proof.* Compactness is obvious. To prove the convexity, we show that, for $x^1 = (x^1_{ia}), x^2 = (x^2_{ia}) \in \mathbb{Q}_{\{\leqq\}\alpha}$ and $\gamma \in (0,1)$, $x = (x_{ia}) \in \mathbb{Q}_{\{\leqq\}\alpha}$ with $x_{ia} = \gamma x^1_{ia} + (1-\gamma)x^2_{ia}, i \in S, a \in A$. Since $x^1, x^2 \in \mathbb{Q}_{\{\leqq\}\alpha}$, there exist $\delta^1 = (\delta^1(i)), \delta^2 = (\delta^2(i))$ such that

$$x^k_i = \beta(i)(1-\delta^k(i)) + \sum_{j\in S, a\in A} x^k_{ja}p_{ji}(a)(1-\delta^k(i)), \quad \text{for } i \in S, k = 1,2. \tag{2.12}$$

Now, define $\delta = (\delta(i))$ as follows:

$$1-\delta(i) = \frac{\gamma x^1_i + (1-\gamma)x^2_i}{\gamma(\beta(i) + \sum_{j,a} x^1_{ja}p_{ji}(a)) + (1-\gamma)(\beta(i) + \sum_{j,a} x^2_{ja}p_{ji}(a)} \tag{2.13}$$

for $i \in S$ where if the denominator is zero, $0 \leqq \delta(i) \leqq 1$ is chosen arbitrary. From (2.12) and (2.13),

it follows that $0 \leqq \delta(i) \leqq 1$ and

$$x_i = \beta(i)(1-\delta(i)) + \sum_{j \in S, a \in A} x_{ja} p_{ji}(a)(1-\delta(i)), \quad i \in S,$$

which implies $x \in \mathbb{Q}_{\{\leqq\}\alpha}$. Also, if $x^k \in \mathbb{Q}_{\{=\}\alpha}$ $(k = 1, 2), x \in \mathbb{Q}_{\{=\}\alpha}$. Thus, $\mathbb{Q}_{\{=\}\alpha}$ is convex.∎

**Theorem 2.3.2.** $\mathbb{Q}_{\{\leqq\}\alpha} = \mathbf{X}^{\beta}_{\{\leqq\}\alpha}(\Pi_S \times \mathcal{S}_S)$.

*Proof.* From Lemma 2.2.1 (ii) and Lemma 2.2.2, the right hand side is clearly contained in the left. To prove the converse, let $x \in \mathbb{Q}_{\{\leqq\}\alpha}$. Then, there exists $\delta = (\delta(i))$ such that $(x, \delta) \in \hat{\mathbb{Q}}_{\{\leqq\}\alpha}$. Define a stationary policy $w$, for any $a \in A$ and $i \in S$, by

$$w(a|i) = \begin{cases} \dfrac{x_{ia}}{x_i}, & \text{if} \quad x_i > 0, \\ \text{any prob. distrib. on } A, & \text{if} \quad x_i = 0 \end{cases}$$

and consider the pair $(w, \tau) \in \Pi_S \times \mathcal{S}_S$ with $\delta = f^{\tau}$. From the definition of $\hat{\mathbb{Q}}_{\{\leqq\}\alpha}$, we have $x_i = \beta(i)(1 - \delta(i)) + \sum_{j \in S} x_j P^{\delta}_{ji}(w)$. Hence, from Lemma 2.2.2, $x_i = x(\beta, w, \tau; i)$. Also, by the definition of $w$, we get

$$x_{ia} = x_i \frac{x_{ia}}{x_i} = x(\beta, w, \tau; i)\frac{x_{ia}}{x_i} = x(\beta, w, \tau; i, a),$$

which implies $x = \{x(\beta, w, \tau; i, a)\} \in \mathbb{X}^{\beta}_{\{\leqq\}\alpha}(\Pi_S \times \mathcal{S}_S)$.∎

From this theorem, we have the following corollary.

**Corollary 2.3.1.** $\mathbb{X}^{\beta}_{\{\leqq\}\alpha}(\Pi_S \times \mathcal{S}_S)$ *is compact and convex.*

Now, define another Mathematical Programming formulation(**MP(II)**) for **COP**:

**MP(II)** : Maximize $\sum_{i \in S, a \in A} c(i, a)x_{ia} + \sum_{i \in S} r(i)y_i$

subject to $(x, \delta) \in \hat{\mathbb{Q}}_{\{\leqq\}\alpha}$,

$$y_i = \beta(i) + \sum_{j \in S, a \in A} x_{ja} p_{ji}(a) - \sum_{a \in A} x_{ia}, \quad i \in S.$$

From Theorem 2.3.1 and 2.3.2, the following corollary easily follows.

**Corollary 2.3.2. COP** *and* **MP(II)** *are equivalent.*

Let $\Pi'_S := \{w \in \Pi_S : w$ requires randomization between two actions in at most one state $\}$, and $\mathcal{S}'_S := \{\tau \in \mathcal{S}_S | f^{\tau}(i) \in \{0, 1\}$ except at most one state $i \in S\}$. For any compact convex set $D$ we denote by ext$(D)$ the set of extreme points of $D$.

**Lemma 2.3.2.**

$$\text{ext}\left(\mathbf{X}^{\beta}_{\{=\}\alpha}(\Pi_S \times \mathcal{S}_S)\right) \subset \{x(\beta, w, \tau) : (w, \tau) \in \Pi'_S \times \mathcal{S}'_S\}. \tag{2.14}$$

*Proof.* By the entire analogy to the proof of Theorem 3.8[4], we can show that

$$\text{ext}\left(\mathbf{X}^{\beta}_{\{=\}\alpha}(\Pi_S \times \mathcal{S}_S)\right) \subset \{x(\beta, w, \tau) : (w, \tau) \in \Pi'_S \times \mathcal{S}_S\}. \tag{2.15}$$

Let $(w, \tau) \in \Pi'_S \times \mathcal{S}_S$. For simplicity, let $\delta = f^{\tau}$. Suppose that there exists $i_1, i_2 \in S (i_1 \neq i_2)$ with $0 < \delta(i_1) < 1, 0 < \delta(i_2) < 1, \mathbb{P}^{w}_{\beta}(X_t = i_1$ for some $t \geq 1) > 0$ and $\mathbb{P}^{w}_{\beta}(X_t = i_2$ for some $t \geq 1) > 0$. We consider $\delta^1 = (\delta^1(i)), \delta^2 = (\delta^2(i))$ satisfying the following (2.16) and (2.17):

$$\begin{cases} \delta^k(i) = \delta(i) \ \text{ if } i \neq i_1, i_2 \quad \text{for each } k = 1, 2, \\ 0 < \delta^1(i_1) < \delta(i_1) < \delta^2(i_1) < 1, \\ 0 < \delta^2(i_2) < \delta(i_2) < \delta^1(i_2) < 1 \end{cases} \tag{2.16}$$

and

$$\begin{cases} \sum_{i \in S} x(\beta, w, \tau^{\delta^1} : i) = \sum_{i \in S} x(\beta, w, \tau^{\delta^2} : i) = \alpha - 1, \\ x(\beta, w, \tau^{\delta^1}) \neq x(\beta, w, \tau^{\delta^2}). \end{cases} \tag{2.17}$$

Note that the existence of such $\delta^k$ $(k = 1, 2)$ is easily shown. For simplicity, let $x^{\delta^1}(i) := x(\beta, w, \tau^{\delta^1}; i)$ and $x^{\delta^2}(i) := x(\beta, w, \tau^{\delta^2}; i)$, $i \in S$. Let $b \in (0, 1)$ be such that

$$1 - \delta(i) = \frac{b x^{\delta^1}(i) + (1-b)x^{\delta^2}(i)}{\beta(i) + \left(\sum_{k \in S}(b x^{\delta^1}(k) + (1-b)x^{\delta^2}(k))(P(w))_{ki}\right)} \tag{2.18}$$

for all $i \in S (i \neq i_2)$.

By the definition of $\delta^1$ and $\delta^2$ we observe that such a $b$ exists. Using this $b \in (0,1)$, we define $\tilde{\delta} = (\tilde{\delta}(i))$ as follows:

$$1 - \tilde{\delta}(i_2) = \frac{bx^{\delta^1}(i_2) + (1-b)x^{\delta^2}(i_2)}{\beta(i_2) + \left(\sum_{k\in S}(bx^{\delta^1}(k) + (1-b)x^{\delta^2}(k))(P(w))_{ki_2}\right)}, \tag{2.19}$$

and $\tilde{\delta}(i) = \delta(i)$ if $i \neq i_2$.

Then, applying Lemma 2.2.2, by (2.18) and (2.19), we get

$$x(\beta, w, \tau^{\tilde{\delta}}) = bx(\beta, w, \tau^{\delta^1}) + (1-b)x(\beta, w, \tau^{\delta^2}). \tag{2.20}$$

By (2.20), $\sum_{i\in S} x(\beta, w, \tau^{\tilde{\delta}}; i) = \alpha - 1$, so that from (2.19), we can assume that $\tilde{\delta} = \delta$. Thus, $x(\beta, w, \tau^{\delta})$ is not an extreme point. The above discussion shows that $\mathrm{ext}(\{x(\beta, w, \tau) : (w, \tau) \in (\Pi'_S \times \mathcal{S}_S)\}) \subset \{x(\beta, w, \tau) : (w, \tau) \in \Pi'_S \times \mathcal{S}'_S\}$. which implies, together with (2.15), that (2.14) holds. ∎

**Theorem 2.3.3.** *For* **COP**, *there exists an optimal pair in* $\Pi'_S \times \mathcal{S}'_S$.

*Proof.* There exists an optimal pair $(w^*, \tau^*) \in \Pi_S \times \mathcal{S}_S$ from Corollary 2.3.1. For $\alpha' := \mathbb{E}_\beta^{w^*}[\tau^*] \leqq \alpha$, $(w^*, \tau^*) \in \mathbb{X}^\beta_{\{=\}\alpha'}(\Pi_S \times \mathcal{S}_S)$. Hence, since the objective function of **MP(II)** is linear, from Lemma 2.3.2 the theorem follows. ∎

**Example.** Here, we give the following numerical example:
$S = \{1,2,3,4\}, A = \{1\}, \alpha = 3, \beta = (0.25, 0.25, 0.25, 0.25)$,

$$(p_{ij}(1)) = \begin{pmatrix} 0.3 & 0.4 & 0.1 & 0.2 \\ 0.4 & 0.1 & 0.2 & 0.3 \\ 0.2 & 0.3 & 0.4 & 0.1 \\ 0.3 & 0.3 & 0.1 & 0.3 \end{pmatrix},$$

$c(1,1) = 0.6, c(2,1) = 0.1, c(3,1) = 0.5, c(4,1) = 0.4, r(1) = 4, r(2) = 3, r(3) = 2, r(4) = 2$.

Letting $x_i = x_{i1}$ $(i \in S)$, the Mathematical Programming formulation (**MP(II)**) for the corresponding **COP** is given as follows:

Maximize $-1.6x_1 - 0.2x_2 + 0.2x_3 + 0.5x_4 + 2.75$
subject to
$x_1 = (0.25 + 0.3x_1 + 0.4x_2 + 0.2x_3 + 0.3x_4)(1 - \delta(1))$,
$x_2 = (0.25 + 0.4x_1 + 0.1x_2 + 0.3x_3 + 0.3x_4)(1 - \delta(2))$,
$x_3 = (0.25 + 0.1x_1 + 0.2x_2 + 0.4x_3 + 0.1x_4)(1 - \delta(3))$,
$x_4 = (0.25 + 0.2x_1 + 0.3x_2 + 0.1x_3 + 0.3x_4)(1 - \delta(4))$,
$x_1 + x_2 + x_3 + x_4 \leqq 2$,
$x_1, x_2, x_3, x_4 \geqq 0, 1 \geqq \delta(1), \delta(2), \delta(3), \delta(4) \geqq 0$.

After a simple calculation, we find that the optimal solution of the above is $x_1^* = 0, x_2^* = 89/156, x_3^* = 113/156, x_4^* = 55/78, \delta^*(1) = 1, \delta^*(2) = 129/574, \delta^*(3) = \delta^*(4) = 0$ and the optimal value is $611/195 (\doteqdot 3.13)$. Note that the value is $75/82 (\doteqdot 3.06)$ for $\delta(1) = \delta(2) = 1$ and $\delta(3) = \delta(4) = 0$.
Thus, by Corollary 2.3.2 and Theorem 2.3.3, the pair $(w^*, \tau^*) \in \Pi'_S \times \mathcal{S}'_S$ with $w^*(i) = 1$ for all $i \in S$ and $f^{\tau^*}(1) = \delta^*(1) = 1, f^{\tau^*}(2) = \delta^*(2) = 129/574, f^{\tau^*}(3) = \delta^*(3) = 0, f^{\tau^*}(4) = \delta^*(4) = 0$ is optimal for the corresponding **COP** and the optimal reward $J(\beta, w^*, \tau^*) = 611/195$.

# 3 Finite MDPs with multiple constraints([19])

## 3.1 Multiple-constrained problem

The aim of this section is to establish a Mathematical Programming method for finite state stopped MDPs with vector-valued terminal reward and multiple running cost constraints. In Section 2, we consider a optimization problem for stopped Markov decision processes with a constrained stopping time. The problem is solved through randomization of stopping times and Mathematical Programming formulation by occupation measures. Here, we consider the vector-valued and multiple constrained case. The optimality is defined by the concept of efficiency, based on a pseudo-order preference relation $\preccurlyeq_K$ induced by a closed convex cone $K$ in $\mathbb{R}^p$, where $\mathbb{R}^p$ denoted the set of real

$p$-dimensional row vectors . Then a Pareto optimization with respect to the pseudo-order $\preccurlyeq_K$ is considered.

Let $K \subset \mathbb{R}^p$ be a nontrivial closed and pointed convex cone (cf. Stoer and Witzgall[36]). We introduce a pseudo-order relation $\preccurlyeq_K$ on $\mathbb{R}^p$ by $x \preccurlyeq_K y$ iff $y - x \in K$. For a nonempty subset $U \subset \mathbb{R}^p$, a point $x \in U$ is called efficient with respect to the order $\preccurlyeq_K$ on $\mathbb{R}^p$ if $x \preccurlyeq_K y$ for some $y \in U$ implies $x = y$. Let $e(U)$ denote the set of all efficient points of $U$ with respect to $\preccurlyeq_K$.

For any $\alpha = (\alpha^1, \ldots, \alpha^k) \in \mathbb{R}^k$ and initial distribution $\beta$ on $S$, let

$$\Lambda^k(\alpha, \beta) := \{(\pi, \tau) \in \Pi \times \mathcal{S} \mid \overline{\mathbb{E}}^\pi_\beta \sum_{t=1}^{\tau-1} c^l(X_t, \Delta_t) \leqq \alpha^l \text{ for } l = 1, \ldots, k\}. \tag{3.1}$$

We shall consider the vector-valued constrained optimization problem (**VCOP**):

**VCOP** : Maximize $\overline{\mathbb{E}}^\pi_\beta \boldsymbol{r}(X_\tau) := (\overline{\mathbb{E}}^\pi_\beta r^1(X_\tau), \ldots, \overline{\mathbb{E}}^\pi_\beta r^p(X_\tau))$
subject to $(\pi, \tau) \in \Lambda^k(\alpha, \beta)$.

A pair $(\pi^*, \tau^*) \in \Lambda^k(\alpha, \beta)$ is called Pareto optimal if

$$\overline{\mathbb{E}}^{\pi^*}_\beta \boldsymbol{r}(X_{\tau^*}) \in e(\{\overline{\mathbb{E}}^\pi_\beta \boldsymbol{r}(X_\tau) | (\pi, \tau) \in \Lambda^k(\alpha, \beta)\}). \tag{3.2}$$

Note that if $c^l \equiv 1$ for $l = 1, 2, \ldots, k$, the running cost constraints are reduced to $\overline{\mathbb{E}}^\pi_\beta \tau \leqq d$, where $d = \min_{1 \leq l \leq k} \alpha^l + 1$, whose case have been studied in Section 2, so that works in this paper are thought of as a generalization of those in Section 2.

Let $K^*$ denote the dual cone of a convex cone $K \subset \mathbb{R}^p$, i.e., $K^* = \{b \in \mathbb{R}^p : \langle b, x \rangle \geq 0 \text{ for all } x \in K\}$ where $\langle \cdot, \cdot \rangle$ means inner product in $\mathbb{R}^p$. The set of interior points of $K^*$ is denoted by $\operatorname{int} K^*$.

The following result is well known (cf. Benson[6]).

**Lemma 3.1.1.** *Let $B \subset \mathbb{R}^p$ be compact and convex set. Then $x \in e(B)$ if and only if there exists $b \in (\operatorname{int} K^*)(b \neq 0)$ such that $\langle b, x \rangle \geq \langle b, y \rangle$ for all $y \in B$.*

## 3.2 Mathematical Programming formulation

Let $\mathbb{R}^{N_1 \times N_2}$ be the set of real $N_1 \times N_2$ matrices. For any subset $U \subset \Pi \times \mathcal{S}$, denote

$$\mathbf{X}^k(U) := \{x(\beta, \pi, \tau; i, a)_{i \in S, a \in A} : (\pi, \tau) \in U \cap \Lambda^k(\alpha, \beta)\}. \tag{3.3}$$

Note that $\mathbf{X}^k(U) \subset \mathbb{R}^{N_1 \times N_2}$.

Here, we define the multi-objective Mathematical Programming problem(**MMP(I)**) related to **VCOP** as follows:

**MMP(I)**:

$$\text{Maximize } \sum_{i \in S} \boldsymbol{r}(i) y(i) := (\sum_{i \in S} r^1(i) y(i), \ldots, \sum_{i \in S} r^p(i) y(i)),$$

subject to $x \in \mathbf{X}^k(\Pi \times \mathcal{S})$, $y \in \mathbb{R}^{N_1}$ and

$$y(i) = \beta(i) + \sum_{j \in S, a \in A} x(j, a) p_{ji}(a) - x(i), \quad i \in S,$$

where $x(i) = \sum_{a \in A} x(i, a)$.

Then, we have the following theorem, which is proved from Lemma 3.1.1 by the use of Theorem 2.2.1.

**Theorem 3.2.1.** **VCOP** *is equivalent to* **MMP(I)**, *i.e., a pair $(\pi^*, \tau^*)$ is Pareto optimal for* **VCOP** *if and only if the corresponding occupation measure $\{x(\beta, \pi^*, \tau^*; i, a)\} \in \mathbf{X}^k(\Pi \times \mathcal{S})$ is Pareto optimal for* **MMP(I)**.

*Proof.* From Lemma 3.1.1, an efficient point for **VCOP** is given by solving the following maximization problem for some $b \in (\operatorname{int} K^*)$:

$$\text{Maximize } \langle b, \overline{\mathbb{E}}^\pi_\beta \boldsymbol{r}(X_\tau) \rangle \quad \text{subject to } (\pi, \tau) \in \Lambda^k(\alpha, \beta). \tag{3.4}$$

Applying Theorem 2.2.1 will complete the proof of Theorem 3.2.1. ∎

## 3.3 Pareto optimal pair

In this section, we present another Mathematical Programming formulation by which **VCOP** is explicitly solved.

To this end, we define several basic sets below. For simplicity, we put $(x_{ia}) = \{x_{ia}\}_{i\in S, a\in A} \in \mathbb{R}^{N_1\times N_2}$ and $\delta = \{\delta(i)\}_{i\in S} \in \mathbb{R}^{N_1}$. For any initial distribution $\beta$ on $S$ and $\alpha = (\alpha^1, \ldots, \alpha^k) \in \mathbb{R}^k$, let

$$\hat{\mathbb{Q}}^k := \left\{ \begin{array}{l} ((x_{ia}), \delta) \in \mathbb{R}^{N_1\times N_2} \times \mathbb{R}^{N_1} : \\ \text{(i)} \sum_{a\in A} x_{ia} = \beta(i)(1-\delta(i)) + \\ \quad \sum_{j\in S, a\in A} x_{ja} p_{ji}(a)(1-\delta(i)), \ (i \in S) \\ \text{(ii)} \ 0 \leqq \delta(i) \leqq 1, \ (i \in S) \\ \text{(iii)} \sum_{i\in S, a\in A} c^l(i,a) x_{ia} \leqq \alpha^l, \\ \qquad\qquad (l = 1, 2, \ldots, k) \\ \text{(iv)} \ x_{ia} \geqq 0, \ (i \in S, a \in A) \end{array} \right\}, \tag{3.5}$$

$$\mathbb{Q}^k := \{(x_{ia}) \in \mathbb{R}^{N_1\times N_2} : ((x_{ia}), \delta) \in \hat{\mathbb{Q}}^k \ \text{ for some } \delta\}. \tag{3.6}$$

We introduce the following assumption.

**Assumption** (*). *For any $w \in \Pi_S$ and $l (1 \leqq l \leqq k)$,*

$$\max_{1\leqq l\leqq k} c^l(i|w) > 0 \ \textit{for each} \ i \in S \tag{3.7}$$

*where $c^l(i|w) = \sum_{a\in A} c^l(i,a) w(a|i)$.*

We have the following theorem, whose proof is similar to ( Theorem 2.3.1, Lemma 2.3.1 and Theorem 2.3.2) and omitted.

**Theorem 3.3.1.** *Suppose that Assumption (*) holds. Then*

(i) $\mathbb{X}^k(\Pi \times \mathcal{S}) = \mathbb{X}^k(\Pi_M \times \mathcal{S}_M) = \mathbb{X}^k(\Pi_S \times \mathcal{S}_S)$.

(ii) $\mathbb{Q}^k = \mathbb{X}^k(\Pi_S \times \mathcal{S}_S)$.

(iii) $\mathbb{Q}^k$ *is compact and convex.*

The following corollary holds clearly from Theorem 3.3.1 and observing (3.6).

**Corollary 3.3.1.** $\mathbb{X}^k(\Pi_S \times \mathcal{S}_S)$ *is compact and convex.*

**Remark.** For any $((x_{ia}), \delta) \in \hat{\mathbb{Q}}^k$, we define a stationary policy $w$ as follows:

For each $a \in A$ and $i \in S$,

$$w(a|i) = \begin{cases} \dfrac{x_{ia}}{x_i}, \ \text{if } x_i > 0, \\ \text{any prob. distrib. on } A, \quad \text{if } x_i = 0, \end{cases} \tag{3.8}$$

where $x_i = \sum_{a\in A} x_{ia}$. Then, $x = (x_i)$ with $x_i = x(\beta, w, \delta; i), i \in S$ is given as a unique solution of (2.3).

Also, (i) and (iii) in (3.5) are rewritten as follows:

$$\begin{cases} \text{(i')} \quad x_i = \beta(i)(1-\delta(i)) + \\ \qquad\qquad \sum_{j\in S} x_j P_{ji}(w)(1-\delta(i)), i \in S \\ \text{(iii')} \quad \sum_{i\in S} c^l(i|w) x_i \leqq \alpha^l, l = 1, 2, \ldots, k \end{cases} \tag{3.9}$$

where $c^l(i|w) = \sum_{a\in A} c^l(i,a) w(a|i)$.

Now, we define another multi-objective Mathematical Programming problem (**MMP(II)**) for **VCOP**:

$$\textbf{MMP(II)}: \ \text{Maximize} \ \sum_{i\in S} r(i) y_i$$

$$\text{subject to} \ (x_{ia}) \in \mathbb{Q}^k,$$

$$y_i = \beta(i) + \sum_{j\in S, a\in A} x_{ja} p_{ji}(a) - \sum_{a\in S} x_{ia}, \ i \in S.$$

Here we get the following corollary which is obviously given from Theorem 3.2.1 and 3.3.1 and Corollary 3.3.1.

**Corollary 3.3.2.** *The following* (i)–(ii) *hold:*

(i) **VCOP** *and* **MMP(II)** *are equivalent.*

(ii) *A Pareto optimal pair exists on* $\Pi_S \times \mathcal{S}_S$.

For any stationary policy $w \in \Pi_S$, let $n(w)$ be the total number of randomization under $w$, that is, $n(w) = \sum_{i\in S}(m(i,w) - 1)$, where $m(i,w)$ is the number of elements in $\{a \in A | w(a|i) > 0\}$. Define $\Pi_S^k := \{w \in \Pi_S : n(w) \leqq k\}$, and $\mathcal{S}_S^k := \{\tau \in \mathcal{S}_S | f^\tau(i) \in \{0,1\}$ except at most $k$ states$\}$. For $(x_{ia}) \in \mathbb{Q}^k$, $\mathcal{I}((x_{ia})) \subset \{1, 2, \ldots k\}$ is defined as follows: $\mathcal{I}((x_{ia})) := \{l \in \{1, 2, \ldots, k\} : \sum_{i\in S, a\in A} c^l(i,a) x_{ia} = \alpha^l\}$. For any $\{l_1, l_2, \ldots, l_h\} \subset \{1, 2, \ldots, k\}$, let $\mathbb{Q}_{\{l_1,\ldots,l_h\}} := \{(x_{ia}) | ((x_{ia}), \delta) \in \hat{\mathbb{Q}}_{\{l_1,l_2,\ldots,l_h\}}$ for some $\delta \in \mathbb{R}^n\}$, where $\hat{\mathbb{Q}}_{\{l_1,\ldots,l_h\}} := \{((x_{ia}), \delta) \in \hat{\mathbb{Q}}^k : \mathcal{I}((x_{ia})) = \{l_1, l_2, \ldots, l_h\}\}$. For any compact convex set $D$ we denote by ext($D$) the set of extreme points of $D$.

Then, we have the following, whose proof is done in Section 3.5.

**Lemma 3.3.1.** *Under Assumption (*), it holds that for any* $\{l_1, \ldots, l_h\} \subset \{1, \ldots, k\}$,

$$\text{ext}\left(\mathbb{Q}_{\{l_1,\ldots,l_h\}}\right) \subset \{x(\beta, w, \delta) : (w, \delta) \in \Pi_S^k \times \mathcal{S}_S^k\}, \tag{3.10}$$

*where $k$ is the number of constraints.*

The existence of a Pareto optimal pair of stationary policy and stopping time requiring randomization in at most $k$ states is given in the following.

**Theorem 3.3.2.** *Suppose Assumption (*) holds. Then a Pareto optimal pair $(\pi^*, \tau^*)$ for* **VCOP** *exists in $\Pi_S^k \times \mathcal{S}_S^k$, that is,*

$$\begin{aligned} &e(\{\mathbb{E}_\beta^\pi r(x_\tau) | (\pi, \tau) \in \Lambda^k(\alpha, \beta)\}) \\ &\subset e(\{\mathbb{E}_\beta^\pi r(x_\delta) | (w, \delta) \in (\Pi_S^k \times \mathcal{S}_S^k) \cap \Lambda^k(\alpha, \beta)\}). \end{aligned} \tag{3.11}$$

**Example 3.1**

Consider the following numerical example with $p = 1$.

$S = \{1, 2, 3, 4\}, A = \{1\}, (\alpha_1, \alpha_2) = (0.5, 0.4), \beta = (0.25, 0.25, 0.25, 0.25)$,

$$(p_{ij}(1)) = \begin{pmatrix} 0.3 & 0.4 & 0.1 & 0.2 \\ 0.4 & 0.1 & 0.2 & 0.3 \\ 0.2 & 0.3 & 0.4 & 0.1 \\ 0.3 & 0.3 & 0.1 & 0.3 \end{pmatrix},$$

$c^1(1,1) = 0.6, c^1(2,1) = 0.1, c^1(3,1) = 0.5, c^1(4,1) = 0.4, c^2(1,1) = 0.6, c^2(2,1) = 0.05, c^2(3,1) = 0.1, c^2(4,1) = 0.8, r(1) = 4, r(2) = 3, r(3) = 2, r(4) = 2$. Letting $x_i = x_{i1}$ $(i \in S)$, the Mathematical Programming problem for the corresponding constrained optimization problem, **(MMP(II))**, is given as follows:

Maximize $-x_1 - 0.1x_2 + 0.7x_3 + 0.9x_4 + 2.75$

subject to

$x_1 = (0.25 + 0.3x_1 + 0.4x_2 + 0.2x_3 + 0.3x_4)(1 - \delta(1))$,

$x_2 = (0.25 + 0.4x_1 + 0.1x_2 + 0.3x_3 + 0.3x_4)(1 - \delta(2))$,

$x_3 = (0.25 + 0.1x_1 + 0.2x_2 + 0.4x_3 + 0.1x_4)(1 - \delta(3))$,

$x_4 = (0.25 + 0.2x_1 + 0.3x_2 + 0.1x_3 + 0.3x_4)(1 - \delta(4))$,

$0.6x_1 + 0.1x_2 + 0.5x_3 + 0.4x_4 \leqq 0.5$,

$0.6x_1 + 0.05x_2 + 0.1x_3 + 0.8x_4 \leqq 0.4$,

$x_i \geqq 0, 0 \leqq \delta(i) \leqq 1, i = 1, 2, 3, 4.$

After a simple calculation, we find the optimal solution of the above is $x_1^* = 0, x_2^* = 26/71$, $x_3^* = 43/71, x_4^* = 57/142$, $\delta^*(1) = 1, \delta^*(2) = 79/209, \delta^*(3) = 0, \delta^*(4) = 33/128$ and the optimal value is $1242/355 (\fallingdotseq 3.49859)$. Note that the value is $285/82 (\fallingdotseq 3.47561)$ for $\delta(1) = \delta(2) = 1$ and $\delta(3) = \delta(4) = 0$.

Thus, by Theorem 3.3.2, the pair $(w^*, \tau^*) \in \Pi_S^2 \times \mathcal{S}_S^2$ with $w^*(i) = 1$ for all $i \in S$ and $f^{\tau^*}(1) = \delta^*(1) = 1, f^{\tau^*}(2) = \delta^*(2) = 79/209, f^{\tau^*}(3) = \delta^*(3) = 0, f^{\tau^*}(4) = \delta^*(4) = 33/128$ is optimal for the corresponding constrained optimization problem and the optimal reward 1242/355. Note that $\tau^* \in \mathcal{S}_S^2$.

## 3.4 Lagrange multiplier approaches

In this section, we define the Lagrangian associated with **VCOP** and the saddle-point statement is given (cf. Kurano et al.[27]). Consequently, by solving a parametric Mathematical Programming problem defined in the sequel, a Pareto optimal pair is obtained.

Let $b = (b_1, \ldots, b_p) \in (\text{int}\, K^*)$. The Lagrangian, $L^b$, associated with **VCOP** is defined as

$$\begin{aligned} &L^b((\pi, \tau), \lambda) := \\ &\sum_{i=1}^{p} b_i \overline{\mathbb{E}}_\beta^\pi(r^i(X_\tau)) + \sum_{l=1}^{k} \lambda_l(\alpha^l - \overline{\mathbb{E}}_\beta^\pi(\sum_{t=1}^{\tau-1} c^l(X_t, \Delta_t))) \end{aligned} \tag{3.12}$$

for any $(\pi, \tau) \in \Pi \times \mathcal{S}$ and $\lambda = (\lambda_1, \ldots, \lambda_k) \in \mathbb{R}_+^k$, where $\mathbb{R}_+^k$ is the positive orthant of $\mathbb{R}^k$.

Hereafter $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_k) \in \mathbb{R}_+^k$ will be written simply by $\lambda \geqq 0$.

For the Lagrangian approach we shall refer to Luenberger[28]. We have the following saddle-point statement, whose proof is similar to (Theorem 2, p.221 in Luenberger[28]) combined with the use of the scalarization technique and omitted.

**Theorem 3.4.1.** (cf. Luenberger[28]) *For some $b \in (\text{int}\, K^*)$, suppose that the Lagrangian $L^b$ has a saddle-point at $(\pi^*, \tau^*) \in \Pi \times \mathcal{S}$ and $\lambda^* \in \mathbb{R}_+^k$, i.e.,*

$$L^b((\pi, \tau), \lambda^*) \leqq L^b((\pi^*, \tau^*), \lambda^*) \leqq L^b((\pi^*, \tau^*), \lambda) \tag{3.13}$$

*for all $(\pi, \tau) \in \Pi \times \mathcal{S}$ and $\lambda \in \mathbb{R}_+^k$. Then, $(\pi^*, \tau^*)$ is a Pareto optimal for* **VCOP**.

In order to have the existence of a saddle-point of the Lagrangian $L^b(b \in (\text{int}\, K^*))$ we introduce the set of $N_1 \times N_2$ matrices as follows:
For $M > 0$, let

$$\mathbb{Q}(M) := \left\{ \begin{array}{l} (x_{ia}) \in \mathbb{R}^{N_1 \times N_2} : \\ \text{(i)} \ \sum_{a \in A} x_{ia} = \beta(i)(1 - \delta(i)) \\ \quad + \sum_{j \in S, a \in A} x_{ja} p_{ji}(a)(1 - \delta(i)) \ (i \in S) \\ \text{(ii)} \ 0 \leqq \delta(i) \leqq 1 \ \ (i \in S) \\ \text{(iii)} \ \sum_{i \in S, a \in A} x_{ia} \leqq M - 1 \\ \text{(iv)} \ \ x_{ia} \geqq 0 \ \ (i \in S, a \in A) \end{array} \right\}. \tag{3.14}$$

Note that $\mathbb{Q}(M)$ is identical with the set of feasible solutions of the Mathematical Programming problem (**MP(II)**) introduced in Section 2 to solve stopped MDPs with a constrained stopping time and condition (iii) of (3.14) means $\overline{\mathbb{E}}^w_\beta \tau^\delta \leqq M$, where $w \in \Pi_S$ is constructed from $(x_{ia})$ through (3.8). Under Assumption $(*)$, it clearly holds that for a sufficient large $M > 0$

$$\mathbb{Q}^k \subset \mathbb{Q}(M). \tag{3.15}$$

Henceforth, $M > 0$ will be fixed such that (3.15) holds.

By using occupation measures defined in Section 2, the Lagrangian $L^b(b \in (\text{int}\, K^*))$ can be rewritten as follows:

$$L^b((x_{ia}), \lambda) := \sum_{i \in S} \sum_{l=1}^{p} b_l r^l(i) y_i + \sum_{l=1}^{k} \lambda_l (\alpha^l - \sum_{j \in S, a \in A} c^l(j, a) x_{ja}) \tag{3.16}$$

$$= \sum_{i \in S, a \in A} (\sum_{j \in S} p_{ij}(a) r^b(j) - r^b(i) - \sum_{l=1}^{k} \lambda_l c^l(i, a)) x_{ia} + \sum_{l=1}^{k} \lambda_l \alpha^l + \sum_{i \in S} r^b(i) \beta(i), \tag{3.17}$$

where $y_i := \beta(i) + \sum_{j \in S, a \in A} x_{ja} p_{ji}(a) - \sum_{a \in A} x_{ia}$ and $r^b(j) := \sum_{l=1}^{k} b_l r^l(j)$, for $(x_{ia}) \in \mathbb{Q}(M)$ and $\lambda \in \mathbb{R}^k_+$.

We need the following condition.

**Assumption** $(**)$. *(Slater condition) There exists* $(x_{ia}) \in \mathbb{Q}(M)$ *such that*

$$\sum_{i \in S, a \in A} c^l(i, a) x_{ia} < \alpha^l \tag{3.18}$$

*for all* $l = 1, \ldots, k$.

Then, applying (Theorem 1, p.217 in Luenberger[28]) we have the following Lemma under Slater condition.

**Lemma 3.4.1.** *Under Assumption* $(*)$ *and* $(**)$, *for any* $b \in (\text{int}\, K^*)$, *the Lagrangian* $L^b$ *has a saddle-point at* $(x^*_{ia}) \in \mathbb{Q}(M)$ *and* $\lambda^* \in \mathbb{R}^k_+$, *i.e.,* $L^b((x_{ia}), \lambda^*) \leqq L^b((x^*_{ia}), \lambda^*) \leqq L^b((x^*_{ia}), \lambda)$ *for all* $(x_{ia}) \in \mathbb{Q}(M), \lambda \in \mathbb{R}^k_+$.

If we construct a stationary policy $w^*$ from $(x^*_{ia}) \in \mathbb{Q}(M)$ in Lemma 3.4.1 through (3.8), $(w^*, \lambda^*)$ satisfies (3.13). Thus, we have the following from Lemma 3.4.1.

**Corollary 3.4.1.** *Under Assumption* $(*)$ *and* $(**)$, *for any* $b \in (\text{int}\, K^*)$, *the Lagrangian* $L^b(\cdot, \cdot)$ *has a saddle-point* $(w^*, \lambda^*) \in \Pi_S \times \mathbb{R}^k_+$.

Applying the results above, we can present a parametric Mathematical Programming approach to obtain a Pareto optimal pair for **VCOP**. For any $b \in (\text{int}\, K^*)$ and $\lambda \in \mathbb{R}^k_+$, let

$$r(i, a|b, \lambda) := \sum_{j \in S} p_{ij}(a) r^b(j) - r^b(i) - \sum_{l=1}^{k} \lambda_l c^l(i, a). \tag{3.19}$$

For $b \in (\text{int}\, K^*)$ and $\lambda \in \mathbb{R}^k_+$, a parametric Mathematical Programming problem $\mathbf{MP}(b, \lambda)$ will be given as follows:

$$\mathbf{MP}(b, \lambda): \ \text{Maximize} \ \sum_{i \in S, a \in A} r(i, a|b, \lambda) x_{ia}$$
$$\text{subject to} \ \ (x_{ia}) \in \mathbb{Q}(M).$$

Then, by using a result in Section 2, for each $\lambda \geqq 0$ we have the optimal value $v(b, \lambda)$ for $\mathbf{MP}(b, \lambda)$. By (3.17) and Lemma 3.4.1, there exists $\lambda^* \in \mathbb{R}^k_+$ such that

$$v(b, \lambda^*) + \sum_{l=1}^{k} \lambda^*_l \alpha^l = \min_{\lambda \geqq 0} (v(b, \lambda) + \sum_{l=1}^{k} \lambda_l \alpha^l). \tag{3.20}$$

From this multiplier $\lambda^*$, we solve $\mathbf{MP}(b, \lambda^*)$. Let $((x^*_{ia}), \delta^*)$ be a solution of $\mathbf{MP}(b, \lambda^*)$. Then, from

the discussion above, $((w^*, \delta^*), \lambda^*)$ is a saddle-point satisfying (3.13), and we can say that $(w^*, \delta^*)$ is a Pareto optimal pair for **VCOP** and the value of $\mathbf{MP}(b, \lambda^*)$ is the expected rewards corresponding the Pareto optimal pair $(w^*, \delta^*)$, where $w^*$ is a stationary policy determined by $x^*_{ia}$ through (3.8).

**Example 3.2**

This is Example 3.1. By solving the equation (3.20) with $b = 1$, we get $\lambda^* = (29/213, 248/213)$ and the value of the saddle-point is 1242/355. In order to obtain a optimal pair for **VCOP**, we solve $\mathbf{MP}(1, \lambda^*)$ and get the optimal pair $(w^*, \tau^*) \in \Pi^2_S \times \mathcal{S}^2_S$ as follows: $w^*(i) = 1$ for all $i \in S$ and $f^{\tau^*}(1) = \delta^*(1) = 1, f^{\tau^*}(2) = \delta^*(2) = 79/209, f^{\tau^*}(3) = \delta^*(3) = 0, f^{\tau^*}(4) = \delta^*(4) = 33/128$ and the corresponding optimal reward 1242/355, which is equal to the numerical results in Example 3.1.

## 3.5 Proof of Lemma 3.3.1

In this section, we prove Lemma 3.3.1.

By argument similar to those used in (Theorem 3.8, p.34, in Altman[4]) we can show that

$$\text{ext}\left(\mathbb{Q}_{\{l_1,\ldots,l_h\}}\right) \subset \{x(\beta, w, \delta) : (w, \delta) \in \Pi^k_S \times \mathcal{S}_S\}. \tag{3.21}$$

Let $(w^*, \delta^*) \in \Pi^k_S \times \mathcal{S}_S$ be such that $x(\beta, w^*, \delta^*) \in \mathbb{Q}_{\{l_1,\ldots,l_h\}}$. Suppose that there exists $j_n (n = 1, \ldots, h+1)$ with

$$0 < \delta^*(j_n) < 1 \quad \text{for} \quad n = 1, 2, \ldots h+1. \tag{3.22}$$

For simplicity, put $x^* = x(\beta, w^*, \delta^*)$ suppressing $\beta, w^*$ and $\delta^*$.

Let $L := \{l_1, l_2, \ldots, l_h\}, \overline{L} := \{1, 2, \ldots, k\} - L, J := \{j_1, j_2, \ldots, j_{h+1}\}$ and $\overline{J} := S - J$. For any row vector $x = (x_1, x_2, \ldots, x_{N_1}) \in \mathbf{R}^n$, we can write $x = (x_J, x_{\overline{J}})$, where $x_J$ and $x_{\overline{J}}$ are subvectors of $x$ and $x_J = \{x_i : i \in J\}$ and $x_{\overline{J}} = \{x_i : i \in \overline{J}\}$. Also, $P^\delta(w^*)$ will be partitioned into submatrices as follows:

$$P^\delta(w^*) = \begin{pmatrix} P^\delta(w^*)_{JJ} & P^\delta(w^*)_{J\overline{J}} \\ P^\delta(w^*)_{\overline{J}J} & P^\delta(w^*)_{\overline{J}\overline{J}} \end{pmatrix},$$

where $P^\delta(w^*)_{JJ} = (P_{ij}(w^*)(1 - \delta(j))), i \in J; j \in J$ and other submatrices are similarly defined.

For simplicity, we write

$$P^\delta(w^*) = \begin{pmatrix} P_1 & P_2 \\ P_3 & Q \end{pmatrix}.$$

Let $c(w^*) = (c_{il}(w^*))$, where $c_{il}(w^*) = c^l(i|w^*)$ for $i \in S$ and $l \in \{1, 2, \ldots, k\}$. $C(w^*)$ will be partitioned as done in the above:

$$C(w^*) = \begin{pmatrix} C_{JL} & C_{J\overline{L}} \\ C_{\overline{J}L} & C_{\overline{J}\overline{L}} \end{pmatrix},$$

suppressing $w^*$.

Here we consider the following inequality system (cf. (3.9)).

$$\begin{aligned} &\text{(i)} && x_J = \beta_J(1 - \delta_J) + x_J P_1 + x_{\overline{J}} P_3, \\ &\text{(ii)} && x_{\overline{J}} = \beta_{\overline{J}}(1 - \delta_{\overline{J}}) + x_J P_2 + x_{\overline{J}} Q, \\ &\text{(iii)} && x_J C_{JL} + x_{\overline{J}} C_{\overline{J}L} = \alpha_L, \\ &\text{(iv)} && x_J C_{J\overline{L}} + x_{\overline{J}} C_{\overline{J}\overline{L}} < \alpha_{\overline{L}}, \end{aligned} \tag{3.23}$$

where $\beta_J(1 - \delta_J) = (\beta(i)(1 - \delta(i)); i \in J), \beta_{\overline{J}}(1 - \delta_{\overline{J}}) = (\beta(i)(1 - \delta(i)); i \in \overline{J})$ and $=$ and $<$ mean componentwise relations.

We note that $x^* = (x^*_J, x^*_{\overline{J}})$ and $\delta^* = (\delta^*_J, \delta^*_{\overline{J}})$ satisfy (3.23) obviously.

From Assumption $(*)$, it clearly holds that $\lim_{n\to\infty} Q^n = \mathbf{0}$, so that $(I - Q)^{-1}$ exists and by (ii) in (3.23) we get

$$x_{\overline{J}} = (\beta_{\overline{J}}(1 - \delta_{\overline{J}}) + x_J P_2)(I - Q)^{-1}, \tag{3.24}$$

where $I$ is an identity matrix with the same dimensions as $Q$.

Also, since (i) in (3.23) includes only $\delta_J$ with respect to $\delta$, it uniquely determines $\delta_J$ if $x_J$ and $\delta_{\overline{J}}$ are given. Thus (i) and (ii) in (3.23) determine uniquely $x_{\overline{J}}$ and $\delta_J$ if $x_J$ and $\delta_{\overline{J}}$ are given. Inserting from (3.24) into (iii) in (3.23), we have that

$$x_J(C_{JL} + P_2(I - Q)^{-1}) = \alpha_L - \beta_{\overline{J}}(1 - \delta_{\overline{J}})(I - Q)^{-1} c_{\overline{J}L}. \tag{3.25}$$

Now, we denote by $\hat{D}$ the set of all pairs $(x_J, \delta_{\overline{J}})$ satisfying (3.23).

Let $D$ be the set of all $x_J, (x_J \geqq 0)$ satisfying (3.25) with $\delta_{\overline{J}} = \delta^*_{\overline{J}}$, that is,

$$D = \{x_J | (x_J, \delta^*_{\overline{J}}) \in \hat{D} \text{ and } x_J \geqq 0\}. \tag{3.26}$$

Observing that (3.25) with $\delta_{\overline{J}} = \delta^*_{\overline{J}}$ has $h$ equations and $h + 1$ unknown elements, we find that $D$ is a

polyhedral convex set with at least one dimension. Since (3.22) means that $x_J^* \in D$ is a relative interior point in $D$, there exists $0 < \gamma < 1$ and $x_J^1, x_J^2 \in D$ with

$$x_J^* = \gamma x_J^1 + (1-\gamma)x_J^2. \tag{3.27}$$

Let $x_{\overline{J}}^1, \delta_{\overline{J}}^1$ and $x_{\overline{J}}^2, \delta_{\overline{J}}^2$ be those determined uniquely thorough (i)–(ii) in (3.23) with $x_J = x_J^1, \delta_{\overline{J}} = \delta_{\overline{J}}^*$ and $x_J = x_J^2, \delta_{\overline{J}} = \delta_{\overline{J}}^*$ respectively. Let $x^1 = (x_J^1, x_{\overline{J}}^1), x^2 = (x_J^2, x_{\overline{J}}^2), \delta^1 = (\delta_J^1, \delta_{\overline{J}}^*)$ and $\delta^2 = (\delta_J^2, \delta_{\overline{J}}^*)$. We can assume that $x^1$ and $x^2$ satisfying (iv) in (3.23) by choosing $x_J^1$ and $x_J^2$ sufficiently near to $x_J^*$. Applying Lemma 2.2.1 we get $x^1 = x(\beta, w^*, \delta^1)$ and $x^2 = x(\beta, w^*, \delta^2)$. Thus, we have that $x(\beta, w^*, \delta^*) = \gamma x(\beta, w^*, \delta^1) + (1-\gamma)x(\beta, w^*, \delta^2)$, which implies $x(\beta, w^*, \delta^*) \notin \mathrm{ext}(\mathbb{Q}_{\{l_1, l_2, \ldots, l_h\}})$. This completes the proof. ∎

# 4 Countable state MDPs with a constraint([20])

## 4.1 Problem formulation

In this section, the optimization problem for a stopped decision process with countable state space is considered. Stopping times $\tau$ are forced to be constrained so that $\mathbb{E}\tau \leqq \alpha$ for some fixed $\alpha > 0$. We introduce a randomized *stationary* stopping time in order to extend the entry time of a stopping region and prove the existence of an optimal constrained pair of stationary policy and stopping time utilizing a Lagrange multiplier approach. In this section, we shall formulate the constrained optimization problem for the countable state space referring to Hordijk [16]. Also, an optimal constrained pair of policy and stopping time is defined. A dynamic system, at times $t = 0, 1, 2, \ldots$, is observed to be in one of a possible number of states. Let $S$ be the countable state space, denoted by $S = \{1, 2, \ldots\}$. We denote by $\mathcal{P}(S)$ the set of all probability vectors on $S$, i.e.,

$$\mathcal{P}(S) := \{p = (p_1, \ldots) | p_i \geqq 0 (i \geqq 1), \sum_{i=1}^{\infty} p_i \leqq 1\}.$$

We allow for breaking down or disappearing of the system with positive probability, so $\sum_{i \in S} p_i \leqq 1$. For each $i \in S$, $\mathcal{P}(i)$ is a subset of $\mathcal{P}(S)$, which is assumed to be given. If at time $t$ the system observed is in state $i$ and the decision maker takes $p(i, \cdot) \in \mathcal{P}(i)$, then the system moves to a new state $j \in S$ selected according to the probability distribution $p(i, \cdot)$. This decision process is then repeated from the new state $j$.

Let $\mathcal{P}$ be the set of all stochastic matrices where i-th row vector $p(i, \cdot) \in \mathcal{P}(i)$. A notion of convergence on $\mathcal{P}$ is given as follows: a sequence $P_n = (p_n(i,j)) \in \mathcal{P}$ converges to $P = (p(i,j)) \in \mathcal{P}$ if $p_n(i,j) \to p(i,j) (n \to \infty)$ for each $i, j \in S$. In this case, we write $\lim_{n \to \infty} P_n = P$. Also, $\mathcal{P}$ with this topology forms metric space (cf. Hordijk[16]). An element of $\mathcal{P}$ is called a transition matrix. The policy $R$ for controlling the system is a sequence of transition matrices, $P_0, P_1, \cdots \in \mathcal{P}$, denoted by $R = (P_0, P_1, \ldots)$, where $P_t$ gives the transition probability at time $t (t \geqq 0)$. Here we confine ourselves to memoryless or Markov policies, which is shown to be sufficient to our optimization problem (cf. Theorem 13.2 in Hordijk[16]). We denote by $\mathcal{R}$ the set of all policies. If the policy takes at all times the same transition matrix, i.e., $P^\infty := (P, P, \ldots), P \in \mathcal{P}$, it is called a stationary policy, denoted simply by $P$ and induces a stationary Markov chain.

The sample space is the product space $\Omega = S^\infty$ such that the projection $X_n$ on the n-th factor S describes the state at time $n$. For each $R \in \mathcal{R}$ and initial state $i \in S$, we can define the measure $\mathbb{P}_{i,R}$ on $\Omega$ in an obvious way. In order to solve our problem described in the sequel, we introduce randomized stopping time (cf. Chow et al.[9], Irle[21] and Kennedy[26]). To this end, enlarging $\Omega$ to $\overline{\Omega} := \Omega \times [0,1]$, let $\mathcal{G}_n = \mathcal{F}_n \times \mathbb{B}_1$, where $\mathcal{F}_n = \sigma(X_0, X_1, \ldots, X_n)$, the $\sigma$-field induced by $\{X_0, X_1, \ldots, X_n\}$, and $\mathbb{B}_1$ is Borel subsets on $[0,1] (n \geqq 0)$ and $\mathcal{G}_\infty = \mathcal{F}_\infty \times \mathbb{B}_1$, where $\mathcal{F}_\infty$ is the smallest $\sigma$-field containing all $\mathcal{F}_n (n \geqq 0)$. Let $N := \{0, 1, 2, \ldots\} \cup \{\infty\}$. We call a map $\tau : \overline{\Omega} \to N$ a (randomized) stopping time with respect to $\mathcal{G} := \{\mathcal{G}_n, n \in N\}$ if $\{\tau = n\} \in \mathcal{G}_n$ for each $n \in N$. The class of stopping times with respect to $\mathcal{G}$ will be denoted by $C(\mathcal{G})$. Let $c : \mathcal{P} \times S \to \mathbb{R}$

and $r : S \to \mathbb{R}$ be running cost and terminal reward functions respectively. For simplicity, we put $c_P(i) := c(P,i)(P \in \mathcal{P}, i \in S)$. Hereafter, we assume that for $P, Q \in \mathcal{P}$ with $p(i,\cdot) = q(i,\cdot)$ $c_P(i) = c_Q(i)$. For any policy $R = (P_0, P_1, \ldots) \in \mathcal{R}$ and $\tau \in C(\mathcal{G})$, we define the expected reward $J_{R,\tau}(i)$ by

$$J_{R,\tau}(i) := \mathbb{E}_{i,R}\left(\sum_{n=0}^{\tau-1} c(X_n) + r(X_\tau)\right), \qquad (4.1)$$

where $\mathbb{E}_{i,R}$ is the expectation with respect to the product measure $\mathbb{P}^*_{i,R} := \mathbb{P}_{i,R} \times \mu$ on $\overline{\Omega}$ and $\mu$ is a Lebesgue measure on $\mathbb{B}_1$. Note that $\tau = \infty$ with positive probability is admissible with zero reward.

A $\tau \in C(\mathcal{G})$ is called randomized stationary if for each $n \geqq 0$,

$$\mathbb{P}^*_{i,R}(\tau = n | X_0, X_1, \ldots, X_{n-1}, X_n = j, \tau \geqq n)$$

is depending only on $j \in S$. In such a case, we can define the set $\{\delta(j), j \in S\}$ by

$$\delta(j) := \mathbb{P}^*_{i,R}(\tau = n | X_0, \ldots, X_{n-1}, X_n = j, \tau \geqq n). \qquad (4.2)$$

Then obviously

$$0 \leqq \delta(j) \leqq 1 \quad \text{for each} \quad j \in S. \qquad (4.3)$$

Conversely, for any set $\{\delta(j), j \in S\}$ satisfying (4.3), we can define a randomized stationary stopping time $\tau$ through (4.2). Such a stopping time is said to be determined by $\{\delta(j)\}$. When $\delta(j) = 0$ or $\delta(j) = 1$ for all $j \in S$, the corresponding stopping time is called simply stationary, which is a entry time of $\Gamma := \{j \in S | \delta(j) = 1\}$, denoted by $\tau_\Gamma$.

Let $\alpha > 0$ be given arbitrarily. Constrained optimal pairs will be defined with respect to a given initial state. So without loss of generality we may assume the initial state is "1". Let

$$\Delta(\mathcal{G}) := \{(R,\tau) \in \mathcal{R} \times C(\mathcal{G}) | \, \mathbb{E}_{1,R}(\tau) \leqq \alpha \text{ and } \mathbb{E}_R(r(X_\tau)) < \infty\},$$

where $\mathbb{E}_R(r(X_\tau))$ denotes the vector with ith component $\mathbb{E}_{i,R}(r(X_\tau))$. In this paper, we will consider the constrained optimization problem:

$$\text{maximize } J_{R,\tau}(1), \text{ subject to } (R,\tau) \in \Delta(\mathcal{G}). \qquad (4.4)$$

The constrained pair $(R^*, \tau^*) \in \Delta(\mathcal{G})$ is called optimal in state $1 \in S$ if

$$J_{R^*,\tau^*}(1) \geqq J_{R,\tau}(1) \qquad (4.5)$$

for all $(R,\tau) \in \Delta(\mathcal{G})$.

We shall use the following.

**Lemma 4.1.1.** (*Generalized dominated convergence theorem cf.* [32, 35])
*Let $P_n, P \in \mathcal{P}$ and $g_n, g, y_n, y$ be vectors with $\lim_{P_n \to P} P_n e = Pe$, $g_n \to g, y_n \to y$ as $n \to \infty$, where $e = (1,1,\ldots)$. If $P_n y_n \to Py$ as $n \to \infty$ and $|g_n| \leqq y_n$ for all $n \geqq 1$, then $P_n g_n \to Pg$ as $n \to \infty$.*

## 4.2 Lagrange formulation for constrained optimization

In this section, the Lagrange multiplier is introduced and the parameterized version of stopped decision process is analyzed.

Introducing the Lagrange multiplier $\lambda \geqq 0$, let

$$c^\lambda_P(i) := c_P(i) - \lambda, \quad i \in S \quad \text{and} \qquad (4.6)$$

$$J^\lambda_{R,\tau}(i) := \mathbb{E}_{i,R}\left(\sum_{n=0}^{\tau-1} c^\lambda(X_n) + r(X_\tau)\right), \quad i \in S \qquad (4.7)$$

for each $(R,\tau) \in \mathcal{R} \times C(\mathcal{G})$. The value function $J^\lambda$ is defined as

$$J^\lambda(i) := \sup_{(R,\tau) \in \mathcal{R} \times C(\mathcal{G})} J^\lambda_{R,\tau}(i). \qquad (4.8)$$

If $J^\lambda(i) = J^\lambda_{R,\tau}(i)$ for all $i \in S$, the pair $(R,\tau)$ is called $\lambda$-optimal.

We need the following assumption.

**Assumption** $(U)$: The following (i)–(iii) are satisfied:

(i) $\mathcal{P}$ is compact and convex,

(ii) $c_P(i) \leqq 0$ for all $P \in \mathcal{P}$ and $i \in S$ and $c_P(i)$ is convex in $P \in \mathcal{P}$ for each $i \in S$

(iii) There exists a vector $u$ with $u \geqq |r|e$ such that

$$e + Pu \leqq u, \text{ and } |c_P|e + Pu \leqq u, \tag{4.9}$$

$$\lim_{N\to\infty} P^N u = 0 \text{ for all } P \in \mathcal{P} \text{ and} \tag{4.10}$$

$$\lim_{P\to P_0} Pu = P_0 u \text{ for all } P_0 \in \mathcal{P}. \tag{4.11}$$

For each $\lambda$, the next theorem holds, under the followings:

$$\mathcal{Q}(\lambda) := \{Q \in \mathcal{P} | \max_{P\in\mathcal{P}}(c_P^\lambda + PJ^\lambda) = c_Q^\lambda + QJ^\lambda\},$$

$$\Gamma(\lambda) := \{i \in S | J^\lambda(i) = r(i)\} \quad \text{and}$$

$$\underline{\Gamma}(\lambda) := \{i \in S | r(i) > \max_{P\in\mathcal{P}}(c_P^\lambda + PJ^\lambda)(i)\}.$$

**Theorem 4.2.1.** (cf. Chap. 3, 4 in Hordijk[16] and Chow et al.[9]) *Suppose that Assumption (U) holds. Then, for any $\lambda \geqq 0$, we have:*

(i) $\sum_{n=0}^{\infty} \mathbb{E}_R |c^\lambda(X_n)| < \infty$ *for all* $R \in \mathcal{R}$.

(ii) $|J^\lambda| \leqq (1+\lambda)u$ *and* $J^\lambda$ *satisfies the following Bellman's optimality equation.*

$$J^\lambda = r \vee \max_{P\in\mathcal{P}}(c_P^\lambda + PJ^\lambda). \tag{4.12}$$

*where* $a \vee b = \max\{a, b\}$ *for real number* $a, b$.

(iii) $P_{i,Q}(\tau_{\underline{\Gamma}(\lambda)} < \infty) = 1$ *for all* $Q \in \mathcal{Q}(\lambda)$ *and a pair* $(Q^\infty, \tau_{\Gamma'})$ *with* $Q \in \mathcal{Q}(\lambda)$ *and* $\underline{\Gamma}(\lambda) \subset \Gamma' \subset \Gamma(\lambda)$ *is* $\lambda$*-optimal in* $i \in S$.

**Corollary 4.2.1.** *Suppose that Assumption (U) holds. Let* $Q(\lambda), \Gamma(\lambda), \underline{\Gamma}(\lambda)$ *be as in Theorem 4.2.1 (iii). Let* $\{\delta(i) : i \in S\}$ *be such that* $0 \leqq \delta(i) \leqq 1$ *and* $\delta(i) = 0$ *if* $i \in \Gamma(\lambda)$, $= 1$ *if* $i \in \underline{\Gamma}(\lambda)$. *Then, for the randomized stopping time* $\tau$ *determined by* $\{\delta(i) : i \in S\}$ *through (4.2), a pair* $(Q^\infty, \tau)$ *with* $Q \in \mathcal{Q}(\lambda)$ *is* $\lambda$*-optimal.*

The next three lemmas are useful in the next section, whose proofs are done by referring to the idea used in (Beutler and Ross[7], and Sennott[33]).

**Lemma 4.2.1.** *For each* $i \in S, J^\lambda(i)$ *is non-increasing and continuous in* $\lambda \geqq 0$.

For some $\lambda$-optimal pair $(Q_\lambda, \tau(\lambda))$ with $Q_\lambda \in \mathcal{Q}(\lambda)$, let

$$V^\lambda(i) := \mathbb{E}_{i,Q_\lambda}\left[\sum_{n=0}^{\tau(\lambda)-1} c(X_n) + r(X_{\tau(\lambda)})\right] \tag{4.13}$$

and

$$K^\lambda(i) := \mathbb{E}_{i,Q_\lambda}\tau(\lambda). \tag{4.14}$$

**Lemma 4.2.2.** *For each* $i \in S$, $K^\lambda(i)$ *and* $V^\lambda(i)$ *are non-increasing in* $\lambda(\lambda \geqq 0)$.

**Lemma 4.2.3.** *It holds that*

(i) *for each* $\lambda \geqq 0, \mathcal{Q}(\lambda)$ *is closed and convex.*

(ii) $\mathcal{Q}(\lambda)$ *is upper semi-continuous in* $\lambda \geqq 0$, *i.e., if* $Q_n \in \mathcal{Q}(\lambda_n), \lambda_n \to \lambda$ *and* $Q_n \to Q$ *as* $n \to \infty$, *then* $Q \in \mathcal{Q}(\lambda)$.

## 4.3 An optimal constrained pair

**Theorem 4.3.1.** *If there exists a non-negative number* $\overline{\lambda}$ *such that*

$$\mathbb{E}_{1,Q_{\overline{\lambda}}}(\tau(\overline{\lambda})) = \alpha \text{ for some } Q_{\overline{\lambda}} \in \mathcal{Q}(\overline{\lambda}), \tag{4.15}$$

$\overline{\lambda}$*-optimal pair* $(Q_{\overline{\lambda}}, \tau(\overline{\lambda}))$ *is an optimal constrained one.*

By Theorem 4.3.1, in order to show the existence of an optimal constrained pair, it is sufficient to prove that there exist the multiplier $\overline{\lambda}$ satisfying (4.15).

To this end, we introduce

$$\gamma := \inf\{\lambda | K^\lambda(1) \leqq \alpha\} \tag{4.16}$$

Since $K^\lambda(1)$ is non-increasing in $\lambda \geqq 0$, $\gamma$ is well-defined in (4.16). Here, we need the following assumption.

**Assumption** (*D*): (*Slater condition* cf. Luenberger[28]) There exists a pair $(R, \tau) \in \mathcal{R} \times C(\mathcal{G})$ such that

$$\mathbb{E}_{1,R}(\tau) < \alpha. \tag{4.17}$$

**Lemma 4.3.1.** *Under Assumption (D),* $\gamma < \infty$.

Let $(\lambda_n)$ and $(\delta_n)$ be any sequences such that

$$\lambda_n > \lambda_{n+1}, \ \delta_n < \delta_{n+1} \ \ (n \geqq 1) \tag{4.18}$$

and $\lim_{n\to\infty} \lambda_n = \lim_{n\to\infty} \delta_n = \gamma$. Then, since $J^\lambda$ is non-increasing in $\lambda$, we have that $\Gamma(\delta_1) \subset \cdots \subset \Gamma(\delta_n) \subset \cdots \subset \Gamma(\lambda_n) \subset \cdots \subset \Gamma(\lambda_1)$. Here, we can prove the following fact.

**Lemma 4.3.2.** *The following holds*:

(i) $\lim_{n\to\infty}\Gamma(\lambda_n)=\Gamma(\gamma)$.

(ii) $\lim_{n\to\infty}\Gamma(\delta_n)\supset\underline{\Gamma}(\gamma)$.

The existence of an optimal constrained pair is given in the following.

**Theorem 4.3.2.** *Suppose that Assumptions* $(U)$ *and* $(D)$ *hold. Then there exists an optimal constrained pair* $(R^*,\tau^*)$ *such that* $R^*$ *is stationary policy and* $\tau^*$ *is a stationary stopping time determined by* $\{\delta(i)\}$ *with* $\delta(i)=1$ *if* $i\notin\underline{\Gamma}(\gamma)$ *and* $\delta(i)=0$ *if* $i\notin\Gamma(\gamma)$ *and requiring randomization in at most one state.*

*Proof.* For any sequences $(\lambda_n),(\delta_n)$ satisfying (4.18), there exist sequences $(\underline{Q}_n),(\overline{Q}_n)$, such that $\overline{Q}_n\in\mathcal{Q}(\lambda_n),(\underline{Q}_n)\in\mathcal{Q}(\delta_n),K^{\delta_n}(1)=\mathbb{E}_{1,\underline{Q}_n}(\tau_{\Gamma(\delta_n)})\geqq\alpha,K^{\lambda_n}(1)=\mathbb{E}_{1,\overline{Q}_n}(\tau_{\Gamma(\lambda_n)})<\alpha$ $(n\geqq 1)$. Noting $\mathcal{P}$ is compact, we can assume that $\underline{Q}_n\to\underline{Q}$ and $\overline{Q}_n\to\overline{Q}$ as $n\to\infty$ for some $\underline{Q}$ and $\overline{Q}\in\mathcal{P}$. By Lemma 4.2.3, $\underline{Q},\overline{Q}\in\mathcal{Q}(\gamma)$. Also, from Assumption $(U)$, $Q^Ne\to 0$ as $N\to\infty$ for all $Q\in\mathcal{P}$, so that, applying Lemma 4.1.1, by Lemma 4.3.2 we get

$$\mathbb{E}_{1,\underline{Q}}(\tau_{\underline{\Gamma}(\gamma)})\geqq\alpha\text{ and}\tag{4.19}$$

$$\mathbb{E}_{1,\overline{Q}}(\tau_{\Gamma(\gamma)})\leqq\alpha.\tag{4.20}$$

If at least one of inequalities (4.19) and (4.20) holds in equality, from Theorem 4.3.1 it follows that there is an optimal constrained pair for state 1.

Suppose that $\mathbb{E}_{1,\underline{Q}}(\tau_{\underline{\Gamma}(\gamma)})>\alpha$ and $\mathbb{E}_{1,\overline{Q}}(\tau_{\Gamma(\gamma)})<\alpha$. We must investigate the following two case. In case that $\mathbb{E}_{1,\underline{Q}}(\tau_{\Gamma(\gamma)})<\alpha$, from Corollary 4.2.1 there exists randomized stopping time $\tau$ determined by $\{\delta(i),i\in S\}$ with $\delta(i)=1$ if $i\in\underline{\Gamma}(\gamma),=0$ if $i\notin\Gamma(\gamma)$ and $0\leqq\delta(i)\leqq 1$ if $\Gamma(\gamma)-\underline{\Gamma}(\gamma)$ and $\mathbb{E}_{1,\underline{Q}}(\tau)=\alpha$, which means from Theorem 4.3.1 that the constrained pair $(\underline{Q}^\infty,\tau)$ is optimal. For this case, obviously $\tau$ can be requiring randomization in at most one state. In case that $\mathbb{E}_{1,\underline{Q}}(\tau_{\Gamma(\gamma)})>\alpha$, noting $\mathbb{E}_{1,\overline{Q}}(\tau_{\Gamma(\gamma)})<\alpha$, there exists $a\in(0,1)$ such that $\mathbb{E}_{1,a\underline{Q}+(1-a)\overline{Q}}(\tau_{\Gamma(\gamma)})=\alpha$. Since $\mathcal{Q}(\gamma)$ is convex, $a\underline{Q}+(1-a)\overline{Q}\in\mathcal{P}$, so that a constrained pair $((a\underline{Q}+(1-a)\overline{Q})^\infty,\tau_{\Gamma(\gamma)})$ is optimal in state 1. ∎

Using the idea of the OLA policy for the usual stopping problem, we can derive some results. For each $\lambda\geqq 0$, let $\Gamma^*(\lambda):=\{i\in S|r(i)\geqq\max_{P\in\mathcal{P}}(c_P^\lambda+Pr)(i)\}$ and $\underline{\Gamma}^*(\lambda):=\{i\in\Gamma^*(\lambda)|r(i)>\max_{P\in\mathcal{P}}(c_P^\lambda+Pr)(i)\}$. Here we introduce an assumption insuring the validity of the OLA stopping time.

**Assumption** $(A_\lambda)$: For any $P=(p(i,j))\in\mathcal{P},p(i,j)=0$ if $i\in\Gamma^*(\lambda)$ and $j\notin\Gamma^*(\lambda)$ or $i\in\underline{\Gamma}^*(\lambda)$ and $j\notin\underline{\Gamma}^*(\lambda)$.

**Corollary 4.3.1.** *Suppose that Assumptions in Theorem 4.3.1 hold and Assumption* $(A_\gamma)$ *holds for* $\gamma$ *as in* (4.16). *Then, we have*:

(i) $\Gamma(\gamma)=\Gamma^*(\gamma)$ *and* $\underline{\Gamma}(\gamma)=\underline{\Gamma}^*(\gamma)$.

(ii) *Let* $\{\overline{J}(i),i\in S\}$ *satisfy that* $\overline{J}(i)=\max_{P\in\mathcal{P}}(c_P^\gamma+P\overline{J})(i)$ *for* $i\in S$ *and* $\overline{J}(i)=r(i)$ *for* $i\in\Gamma^*(\gamma)$. *Then, for the initial state "1",* $\overline{J}(1)=\sup_{(R,\tau)\in\Delta(\mathcal{G})}J_{R,\tau}(1)$.

**Example.** Here we give a simple example for a Markov deteriorating system with state space $S=\{1,2,\ldots\}$. This system is formulated as follows:

(i) $\mathcal{P}\subset\{P=(p(i,j))|\sum_{j\in S}p(i,j)=\beta,p(i,j)\geqq 0\text{ for }i,j\in S\}$ for some $\beta(0<\beta<1)$ and $\mathcal{P}$ is convex and compact.

(ii) For any $P=(p(i,j))\in\mathcal{P},p(i,j)=0$ if $i>j$.

(iii) $c_P(i)=-c$ for some $c>0$.

(iv) The reward function $r$ on $S$ has a property that for each $P\in\mathcal{P},(Pr-r)(i)$ is non-increasing in $i\in S$.

Under these assumptions, we observe that Assumptions $(U)$ and $(D)$ hold. Also, by simple calculation we find that for $\lambda\geqq 0$ there exists non-negative integer $i_\lambda\leqq\underline{i}_\lambda$ such that $\Gamma^*(\lambda)=[i_\lambda,\infty)$ and $\underline{\Gamma}^*(\lambda)=[\underline{i}_\lambda,\infty)$, so that Assumption $(A_\lambda)$ hold for all $\lambda\geqq 0$. Thus, for any $\alpha>0$, from Corollary 4.3.1 we know that there exists an optimal constrained pair for this system.

## 参考文献


[1] E. Altman. Denumerable constrained Markov decision process and finite approximations. *Math. Oper. Res.*, 19:169–191, 1994.

[2] E. Altman. Constrained Markov decision processes with total cost criteria: occupation measures and primal lp. *Math. Methods Oper. Res.*, 43:45–72, 1996.

[3] E. Altman. Constrained Markov decision processes with total cost criteria: Lagrangian approach and dual linear program. *Math. Methods Oper. Res.*, 48(3):387–417, 1998.

[4] E. Altman. *Constrained Markov Decision Processes.* Chapmann & Hall/CRC, 1999.

[5] D. Assaf and E. Samuel-Cahn. Optimal multivariate stopping rules. *J. Appl. Probab.*, 35:693–706, 1998.

[6] H. P. Benson. An improved definition of proper efficiency for vector maximization with respect to cones. *J. Math. Anal. Appl.*, 71(1):232–241, 1979.

[7] F. J. Beutler and K. W. Ross. Optimal policies for controlled Markov chains with a constraint. *J. Math. Anal. Appl.*, 112:236–252, 1985.

[8] V. S. Borkar. *Topics in controlled Markov chains.* Pitman Reserch Notes in Mathematics Series, 240. Longman Scientific & Technical, Harlow; John Wiley & Sons, Inc., New York, 1991.

[9] Y. S. Chow, H. Robbins, and D. Siegmund. *Great expectations: the theory of optimal stopping.* Houghton Mifflin Co., Boston, Mass., 1976.

[10] C. Derman. *Finite state Markovian decision processes.* Academic Press, New York, 1970.

[11] C. Derman and M. Klein. Some remarks on finite horizon Markovian decision models. *Operations Res.*, 13:272–278, 1965.

[12] E. B. Frid. On optimal strategies in control problems with constraints. *Theory of Probability and its Applications*, 17:188–192, 1972.

[13] N. Furukawa. Functional equations and Markov potential theory in stopped decision processes. *Mem. Fac. Sci. Kyushu Univ. Ser. A*, 29(2):329–347, 1975.

[14] N. Furukawa. Characterization of optimal policies in vector-valued Markovian decision processes. *Math. Oper. Res.*, 5(2):271–279, 1980.

[15] N. Furukawa and S. Iwamoto. Stopped decision processes on complete separable metric spaces. *J. Math. Anal. Appl.*, 31:615–658, 1970.

[16] A. Hordijk. *Dynamic programming and Markov potential theory.* Mathematical Centre Tracts, No. 51. Mathematisch Centrum, Amsterdam, 1974.

[17] A. Hordijk and L. C. M. Kallenberg. Constrained undiscounted stochastic dynamic programming. *Math. Oper. Res.*, 9:276–289, 1984.

[18] M. Horiguchi. Markov decision processes with a stopping time constraint. *Math. Methods Oper. Res.*, 53:279–295, 2001.

[19] M. Horiguchi. Stopped markov decision processes with multiple constraints. to appear in Math. Methods Oper. Res., Vol. 54(2001) issue 3.

[20] M. Horiguchi, M. Kurano, and M. Yasuda. Markov decision process with constrained stopping times. In *Proceedings of 39th IEEE Conference on Decision and Control. (CDC2000)*, volume 1.

[21] A. Irle. Minimax results and randomization for certain stochastic games. *Minimax Theory and Applications(Erice, 1996)*, pages 91–103, 1998.

[22] S. Iwamoto. Stopped decision processes on compact metric spaces. *Bull. Math. Statist.*, 14(1-2):51–60, 1970.

[23] Y. Kadota, M. Kurano, and M. Yasuda. Utility-optimal stopping in a denumerable Markov chain. *Bull. Inform. Cybernet.*, 28:15–21, 1996.

[24] Y. Kadota, M. Kurano, and M. Yasuda. On the general utility of discounted Markov decision processes. *Int. Trans. Opl. Res.*, 5:27–34, 1998.

[25] L. C. M. Kallenberg. *Linear programming and finite Markovian control problems.* Mathematisch Centrum, Amsterdam, 1983.

[26] D. P. Kennedy. On a constrained optimal stopping problem. *J. Appl. Probab.*, 19:631–641, 1982.

[27] M. Kurano, J.-I. Nakagami, and Y. Huang. Constrained Markov decision processes with compact state and action spaces: the average case. *Optimization*, 48(2):255–269, 2000.

[28] D. G. Luenberger. *Optimization by vector space methods.* John Wiley & Sons, Inc., New York-London-Sydney, 1969.

[29] D. C. Nachman. Optimal stopping with a horizon constraint. *Math. Oper. Res.*, 5:126–134, 1980.

[30] U. Rieder. On stopped decision processes with discrete time parameter. *Stochastic Processes Appl.*, 3(4):365–383, 1975.

[31] S. M. Ross. *Applied probability Models with Optimization Applications.* Holden-Day, 1970.

[32] H. L. Royden. *Real Analysis.* Macmillan, New York, 2d ed. edition, 1968.

[33] L. I. Sennott. Constrained discounted Markov decision chains. *Probability in the Engineering and Informational Sciences*, 5:463–475, 1991.

[34] L. I. Sennott. Constrained average cost Markov decision chains. *Probability in the Engineering and Informational Sciences*, 7:69–83, 1993.

[35] L. I. Sennott. *Stochastic Dynamic Programming and the Control of Queuing Systems.* John Wiley and Sons, Inc., 1999.

[36] J. Stoer and C. Witzgall. *Convexity and optimization in finite dimensions. I.* Springer-Verlag, New York, 1970.

[37] K. Wakuta. Optimal stationary policies in the vector-valued Markov decision process. *Stochastic Process. Appl.*, 42(1):149–156, 1992.